REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	28.Jul.99	THESIS

4. TITLE AND SUBTITLE
DISPLAY OF PREDICTOR RELIABILITY ON A COCKPIT DISPLAY OF TRAFFIC INFORMATION

5. FUNDING NUMBERS

6. AUTHOR(S)
2D LT GEMPLER KEITH S

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITY OF ILLINOIS AT URBANA

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
THE DEPARTMENT OF THE AIR FORCE
AFIT/CIA, BLDG 125
2950 P STREET
WPAFB OH 45433

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
FY99-178

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION AVAILABILITY STATEMENT
Unlimited distribution
In Accordance With AFI 35-205/AFIT Sup 1

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

14. SUBJECT TERMS	15. NUMBER OF PAGES	16. PRICE CODE
	79	

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT

DISPLAY OF PREDICTOR RELIABILITY ON A COCKPIT DISPLAY OF TRAFFIC INFORMATION

BY

KEITH STEWART GEMPLER

B.S. UNITED STATES AIR FORCE ACADEMY, 1997

THESIS

Submitted in partial fulfillment of the requirements
For the degree of Master of Science in Psychology
In the Graduate College of the
University of Illinois at Urbana-Champaign, 1998

Urbana, Illinois

Table of Contents

1.0 Introduction

To improve the availability of information to the pilot concerning other traffic, the concept of a Cockpit Display of Traffic Information (CDTI) has been developed through efforts by NASA. These displays make information about the pilot's own aircraft and others in the flying environment visible, enabling pilots see potential conflicts and avoid them with the most effective maneuvering.

These displays support the challenge of free flight (RTCA 1995), where the pilot becomes more autonomous in deciding exact routing of his aircraft between destinations. With this autonomy from Air Traffic Control, comes an increase in requirements for the pilot to be aware of the position of both his own aircraft and other traffic that may pose a conflict (Wickens, Mavor, Parasuraman, & McGee, 1998). Therefore, information about ownship and othership's current and future positions must be displayed so the pilot can choose a course, speed, and altitude that will maintain safe separation from other aircraft. To increase the efficiency of maneuvers (saving costs in terms of fuel and delays) the pilot will need to make maneuvering decisions based on predicted aircraft separation well in advance of a possible conflict (Wickens, et al, 1998). The development of this CDTI system has raised several psychological issues, many of which have already been investigated.

Previous free-flight oriented CDTI studies have examined the basic structure of the CDTI, including ownship and othership symbology, design of predictor and threat display features, and the dimensionality of the display

(Merwin and Wickens, 1996; Johnson, et al, 1997; Ellis, McGreevy, and Hitchcock, 1987; Hart and Wempe, 1979; Kriefeldt, 1980; Palmer, 1983). Thus far, much of the evidence points to a coplanar (both lateral and vertical) display to be most effective for maneuvering in three physical dimensions and time as compared to plan-view and three dimensional displays (Merwin and Wickens, 1996; Merwin, O'Brien, and Wickens, 1997). One important aspect of CDTI symbology has been the use of predictor lines on both ownship and other traffic which show the estimated future location of the aircraft over a varying span of time (Wickens and Morphew, 1997). While the predictor line has become critical in realizing the potential benefits of CDTI, the impact of the reliability of that predictor has yet to be examined. The focus of this study is on trust calibration and its resultant performance impact in the face of unreliable predictive displays. In the following review, we will first examine basic tracking, which is representative of flight control, and reveal the limitations of human prediction in tracking tasks. Secondly, we will see how research shows that predictive displays can help to overcome human limitations at tracking. Thirdly, we will describe predictive displays as a form of automation, which inherently can not be perfectly reliable. Finally, we will examine how the human operator behaves when faced with less than perfectly reliable information, and specifically in the case of the CDTI, we will attempt to evaluate a display that contributes to the proper calibration of trust in a less than perfectly reliable predictor.

1.1 General Tracking

Flying an aircraft on a specified path is a type of tracking task (Wickens and Carswell, 1997). The pilot performing the tracking task ideally will minimize error (the sum of deviations between output and command input), instability, control activity, and workload (Wickens, 1986). However, each of these criteria affects the others, and based on their relative importance to the task, may be traded for better performance in another area. For example, one could minimize error while increasing control activity, or decrease workload at a cost to error. This behavior can be modeled mathematically, given some constraints, as in the Optimal Control Model (Levison, 1982) (Figure 1).

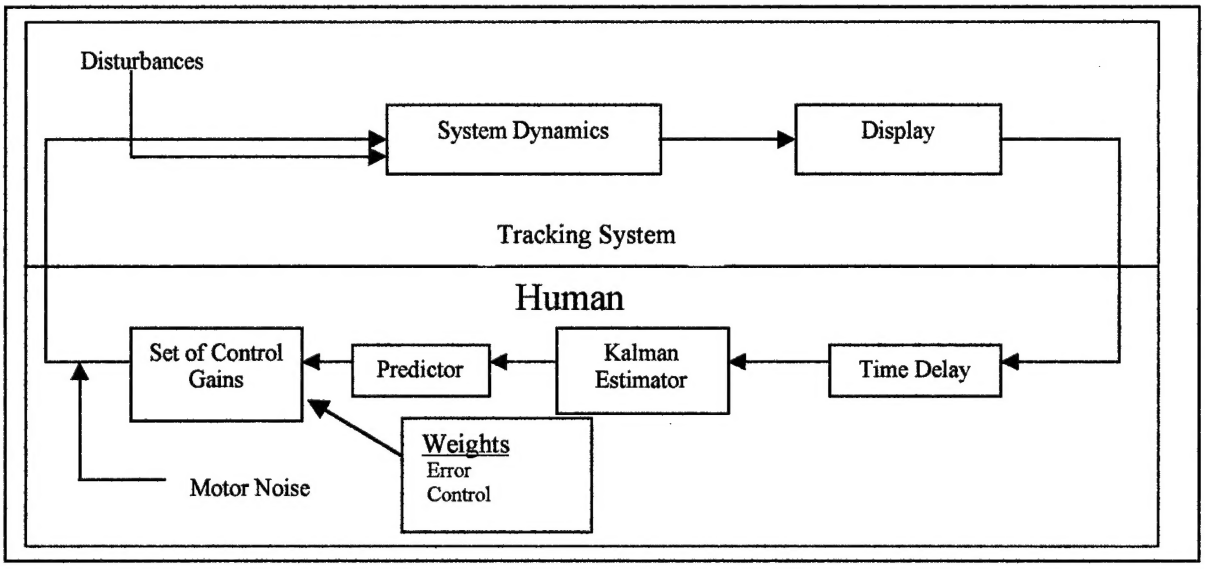


Figure 1: Levison (1982) Optimal Control Model

In this model, the human attempts to minimize display error, control displacement, and control velocity, each of which is weighted by its relative importance to the operator.

The Optimal Control Model consists of disturbances to a system, which are displayed to the operator but perturbed by some observation noise, like that incorporated in signal detection theory, and a time delay caused by internal human processing. The Kalman estimator and predictor serve to represent the pilot's estimate of current system state and the expectation of some future state of the system, respectively. The performance criteria (minimize error, minimize control inputs, etc) are dealt with through a set of control gains, and final control movements are carried out with some associated motor noise. The Optimal Control Model captures several limitations of the human when performing a tracking task. Specifically, to overcome the time lag between signal input and control input, the human must anticipate where the signal is going to be in the future (seen in the predictor portion of the model). This would demand that the human "see into the future" as to where the signal will be several seconds ahead. Humans, however, are relatively poor at making the prediction as to where that signal will be in the future (Wickens 1986).

1.2 Prediction

In the free-flight environment, pilots will be required to predict possible future conflicts in the process of determining the proper course and altitude to fly enroute. However, humans are relatively poor at predicting future position, based on current and historical velocity. "This performance deficit is present because prediction normally requires some sort of mental computation necessary to extrapolate the future from the present and past and may possibly

require that the future estimates be stored in working memory until they are used" (Wickens, 1986). The performance deficit is evident in the use of CDTIs in lab settings, but has been shown to be reduced by providing predictive information on the display, thus changing a cognitive process into a perceptual one. The display makes perceptually visible, quantities that would otherwise need to be cognitively derived (Wickens and Morphew, 1997). The imperfect cognitive process of prediction can be improved through this automated prediction display.

Predictive displays indicate the future state of the system, based on certain assumptions, for example: an algorithm using present velocity values, a filed or programmed flight plan, information resident in the flight management system, or active queries of the pilots involved (Wickens, Mavor, Parasuraman, & McGee, 1998; Gallaher, Hunt, and Willeges, 1997). As we have described above, predictive displays have been shown to be useful in general (Jensen, 1981; Kelly, 1968;), and in the setting of the CDTI in particular (Johnson, et al, 1997; Palmer, Jago, DuBord, 1981; Wickens and Morphew, 1997). Preview, on the other hand, shows the future signal input, and is not inferred from the environment by automation or the human, but can be directly seen. For example, a motor vehicle operator would have a difficult job of driving a car down a winding road (a two dimensional tracking task) if he could see no more than a few feet in front of the vehicle. But, by seeing the upcoming road, the driver is given a preview of the road (signal) and can more accurately anticipate and compensate for bends in the road (changes in the signal).

Furthermore, by perceiving the heading of the vehicle, the driver can infer the future lateral deviation of the vehicle from the road, thus predicting that he will drive off of the road ahead with his current course. Thus, preview is directly observed, but prediction must be inferred, either by the human or by some automated process that displays the prediction. In either case, prediction is inherently less than perfectly reliable.

Poulton (1957) describes this same situation in an experiment where a subject is "given a ballpoint pen which could be moved against a transparent bar. The subject had to keep the point on a curved line drawn on a paper tape. The tape moved at a constant speed at right angles to the bar, and the curved line on it could be seen (previewed) for varying distances ahead of the bar." Subjects were able to follow the track of the curved line more accurately by being given a preview of the upcoming curves in the lines, and predicting whether they would remain on the line given their current position on the bar, just as the driver can track the road easily with sufficient preview of the road ahead and inferring his future position relative to the edge of the road. The distance ahead, or time in the future that the driver infers his future position would be the "span of prediction". If the driver's span of prediction was very short (say a few feet ahead of the vehicle) then his prediction would be more accurate than a longer span of prediction (say several hundred yards ahead of the vehicle) due to inaccuracies inherent in prediction that are perpetuated and magnified at each increase in prediction span.

Jensen (1981) conducted an experiment involving professional pilots tracking an aircraft along curved landing approaches in a Gat-2 flight simulator using predictive and quickened perspective flight displays. He found that lateral error on the curved approach task was reduced with increasing accuracy of prediction using higher orders of computation.

Jago and Palmer (1982) evaluated four types of predictors on a CDTI using 16 airline pilots with no CDTI experience who were required to judge the future position of an intruder aircraft in relationship to their own. They did not interact with any control, but merely responded by keystroke whether the intruder would pass in front of or behind their aircraft after viewing the developing scenario on the display for 8 seconds. Evidence showed that the addition of all tested predictors reduced the error rate, with the greatest reduction coming from the addition of a predictor that included turn rate information on both the ownship and intruder aircraft.

In a study by Johnson, Battiste, Delzell, Holland, Belcher, and Jordan (1997), the authors developed and demonstrated a prototype free flight cockpit display of traffic information called a CSD (Cockpit Situational Display). A baseline display included only a plan view of traffic with no speed and heading changes, and without highlighting potential conflicts (or lowlighting aircraft without conflict potential). In their "enhanced display", they attempted to provide sufficient predictive information to help pilots visualize future conflicts thus reducing the prediction deficit of the human operators. The enhanced display included a temporal predictor that showed the

ownship and othership position between 0 and 10 minutes into the future. The predictor was based upon the assumption that the aircraft would continue along its present track at its present speed. The added predictive information for the pilot to use in maneuver planning, incorporated in the enhanced display, did allow more proactive responses by the participating crews, as well as increased separation distance relative to the maneuvers taken using only the baseline "present position" display. The researchers called for more study to examine how to represent the certainty of predictions on a predictive display because they did not vary the reliability of the predictor. This study augments the information provided by Jago and Palmer (1982) by introducing the more demanding and interactive task of maneuvering around other conflicting traffic.

Barhydt and Hansman (1997) have experimented with adding CDTI predictive displays to Traffic Alert and Collision Avoidance Systems (TCAS) with much success. An experiment was run on a part task flight simulator using eight airline pilots as participants. Predictive information was displayed on a CDTI using four types of predictors, based on several potential sources of information. In one configuration, no predictive information was given using only a current TCAS display. The next display incorporated information on current heading. The next addition was commanded (intended) heading and altitude put into the autopilot of the intruder aircraft. The final display added FMS path information to the display, depicting the intruder's programmed strategic route plan both vertically and horizontally. In addition, the predictive

displays had a conflict probe, which indicated where in the future, minimum separation (2nm lateral or 500ft vertical) would be violated.

Barhydt and Hansman found that all three predictive displays reduced the number of separation violations and lead to earlier responses to developing conflicts relative to the baseline display, and that pilots preferred any of the predictive display types to the non-predictive TCAS display. Fewer separation violations occurred when pilots initiated maneuvers earlier in the scenario, suggesting that any display should encourage pilots to make small changes early in a scenario.

The authors assumed that the appropriate source of predictive information would be available for the intruder aircraft and that the information would perfectly predict the intruders' future position. In general, their data showed that any prediction gained from these information sources which allowed the pilots to initiate their maneuvers earlier lead to fewer conflicts.

Wickens and Morphew (1997) found that a predictive display not only increased performance, but also reduced workload over a non-predictive display. Wickens and Morphew had pilots fly a series of traffic avoidance maneuvers in a part task simulator using three types of CDTI displays. A baseline display provided current position and predictive information regarding ownship and current position information for a single intruder aircraft. The second display type added a predictor for the intruder aircraft as well. The final display included both predictors and added a "threat vector", which, like

Barhydt and Hansman's (1997) conflict probe, indicated the direction of ownship from the intruder at the predicted point of closest contact. Both of the predictor displays lead to fewer conflicts and fewer predicted conflicts. Both enhanced displays also reduced pilot workload as measured by the NASA-TLX scale, and the threat vector display did so to a greater extent than the intruder predictor display. This study expands further upon the previous CDTI prediction research by adding secondary tasks of navigation, as well as an experimental workload task. The workload task consisted of monitoring an area above the CDTI display for indicators that appeared at unpredictable intervals and then hitting a key when the indicator was detected. This task simulated the monitoring of an out-of-the-cockpit view. Response times to and accuracy of detecting the indicators were recorded. No effects on the secondary task were found.

The studies of CDTI prediction have shown that in general, prediction increases overall performance, perhaps by reducing the workload required by internally generating a prediction of future position. However, an important factor that has been omitted from all previous work on CDTI has been that concerning the reliability of the automation which makes inferences about the future to incorporate into the predictor display. The addition of automation into the cockpit brings with it issues of trust and automation reliability which are discussed in the following sections.

1.3 Prediction as Automation

To overcome the inadequacies of human operators in predicting future location of a target, a level of automation can be introduced to perform the extrapolation task which is necessary for a predictor display. Prediction on the CDTI represents a form of automation as defined by Parasuraman and Riley (1997), as "the execution by a machine agent (usually a computer) of a function that was previously carried out by a human." The human task of predicting future positions of aircraft on the CDTI display would now be carried out (much more accurately and with demonstrated benefit) by the automation.

Several of a pilot's decisions and actions will be based upon the automated prediction in the CDTI, including:

Decisions:

1. Whether a loss of separation (LOS) is likely to occur based on a threshold of likeliness in the pilot's mind
    a. based on regulations (e.g. the pilot knows the intruder will slow to 250 kts below 10,000 feet, while the predictor may be based on a higher current velocity)
    b. based on past experience
    c. based on predictor reliability
2. Which maneuvers, if any, will avoid this impact
3. Where in space and time this maneuver should take place

Actions:

1. Change in route, altitude, or speed

2. Communications with other aircraft or supervisory agency for coordination in maneuvering

An important issue of automated prediction is that there is an inherent unreliability in predicting trajectories, so no matter how accurate information is about present position and velocity, an estimate of future position is less than 100% reliable. Small changes in variables occur randomly over time. For example, a pilot may not track the exact heading displayed on the Flight Management System. Errors in engine sensing gauges may change speeds. Changing climactic conditions, from wind direction changes to temperature and air density changes, can impact future position, as can a pilot's own decision to fly a course different from that which was intended and/or programmed into the autopilot. None of these variables can be foreseen with perfect accuracy, especially that of human behavior and weather, and thus prediction is inherently unreliable, and accuracy becomes less perfect as the span of prediction increases. This is a potential problem, because the CDTI scenario requires decisions to be made at time 'x' that will impact the scenario several minutes into the future, when prediction becomes inherently more unreliable. The inherent decreasing reliability in prediction is further confounded because any pilot decision or action based on and affecting the predicted situation, for example three minutes ahead, also impacts the scenario ten minutes ahead and

so forth. However, if the pilot's planning only considers prediction 3 minutes ahead, and completely disregards further prediction, then the unreliable information beyond the three minute mark would have no impact on pilot behavior.

Since predictors can not be perfectly reliable, an understanding of the consequences of unreliability on human performance is of considerable importance. We discuss these issues in the following section.

1.4 Human Performance with Unreliable Automation

The goal of the introduction of automation into complex systems has been to reduce the probability of accidents and inefficiencies caused by "Human Error". However, that introduction of automation into complex systems, like the airline cockpit, has had an impact on error not seen in less automated systems. Much research has been directed at the issue of complacency, or over-reliance on automation (e.g. see Parasuraman and Riley, 1997). Two examples which illustrate the safety impact of complacency in the modern cockpit are Eastern Airlines flight 401 (National Transportation Safety Board, 1973), and an Airbus A330 test flight accident (Sparaco,1994). In the first case, the aircraft on autopilot began a slow descent into terrain in the Florida everglades while the crew was attending to the disassembly of a warning light. Despite cockpit indications of the descent, the pilots failed to notice the automation failure and resulting descent until too late to recover the aircraft. In the second case, an Airbus A330 crashed during a test flight which

was to evaluate the autopilot in an engine-out hydraulic failure on takeoff. The crew did not assume manual control of the aircraft in time to recover from an autopilot error, and all aboard were killed in the resulting impact. Both of these scenarios highlight the safety problems involved with complacency using less than perfectly reliable automation.

Complacency is one type of failure to calibrate perceived reliability with actual system reliability. Figure 2 shows how actual reliability and perceived reliability should be in a linear relationship. However, when perceived reliability of the system is greater than actual reliability, over-trust and complacency occur. While on the flip side, perceived reliability that is lower than actual reliability leads to under-trust and inefficiency. The focus of this study will be on the region of complacency.

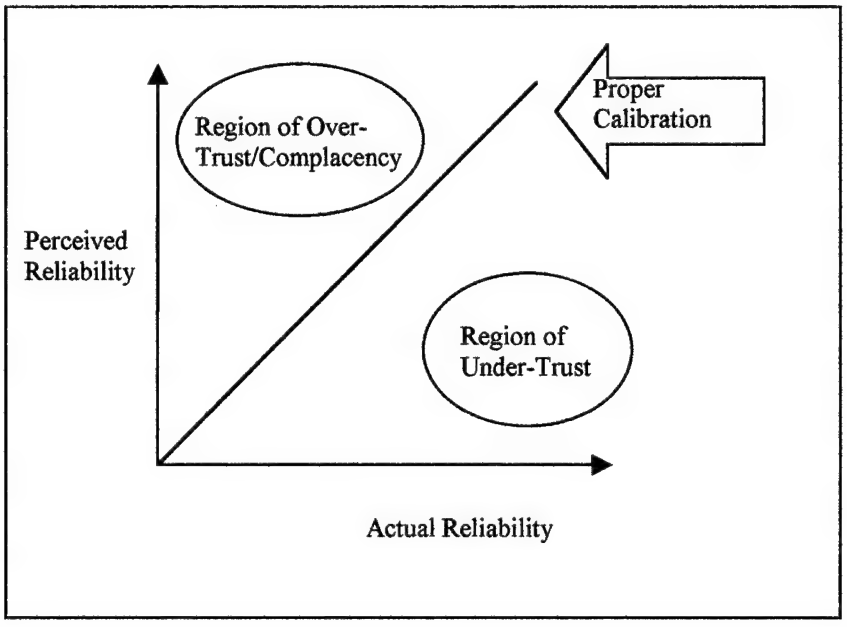


Figure 2: Reliability calibration

In both of the accident cases cited at the beginning of this section, we can see that the operators needed to integrate information from several sources to form a likelihood of possible states of the world (Wickens, Pringle, and Merlo, in preparation). The difference between the two possible states of the world was that either the automation was performing as required, or that it was not. Each possible state of the world has a corresponding set of required actions. In the case of properly performing automation, the pilots could concentrate on other activities, like replacing the light bulb in the Eastern Airlines flight 401 example. In the case of failed automation, manual control would need to be implemented. Based upon the pilots' perceived reliability of the automation, neither crew put much weighting in the information sources (monitors of altitude, descent rate, the visual clues of descent or ground rush) that disagreed with the hypothesis that the automation was not working properly. Thus, they failed to properly integrate the disagreeing sources of information into their hypothesis that the automation was working properly. Their over-trust bias, in turn, lead them to allocate insufficient resources to planning for the alternative hypothesis (Wickens, Pringle, and Merlo, 1998), that the automation had failed and they would need to recover the aircraft.

The consequences of complacency are obvious in the accident examples given in the beginning of this section. Response strategies to failures are less attended to, thus fewer resources are put into developing plans for failure recovery. In the Airbus A330 example, if the pilots had taken manual control of the aircraft 4 seconds earlier in the flight, they could have recovered the aircraft. They did not plan for failure, because their perceived reliability of the automation was higher than its actual reliability, thus falling into the region of

complacency shown in figure 2. In the case of the Eastern Airlines accident, the over-trust lead the pilots to entirely ignore monitoring of altitude and the automation that they trusted to fly the aircraft while they were occupied.

Lee, Moray, and Muir have performed several detailed studies of changes in operators' trust in automation with and without faults (Muir 1987; Lee and Moray 1992, 1994; Muir and Moray, 1996). Generally these studies have entailed the use of a simulated semi-automated processing plant. The term semi-automation refers to the fact that the subsystems of the plant could be run by automation or manual control. Thus, there were two possible states of the world, properly functioning automation or failed automation. There were also two corresponding actions, either using manual control in failed automation or using automatic control. Operators were observed under varying circumstances of automation failures. The performance measurements of these experiments were changes in use of automatic controllers of the subsystems of the plant. Subjective ratings of trust and performance on the operation of the plant were also measured.

While results indicate that individual biases make up a large percentage of the variance in control strategies (Lee and Moray 1994), trust in automation also stems from several other factors, including the newness of the technology, an understanding of the technology, and experience with the technology over time (Lee and Moray 1992). Using theories of trust developed by those studying relationships between humans, the researchers predicted that trust

would undergo predictable changes as a result of failures and other experiences with the automation.

Muir and Moray (1996) found that any sign of incompetence in the automation significantly reduced trust, even when the incompetence had no visible effect on the performance of the automated system. Operators' subjective ratings of trust in the automated system were based almost solely upon their perception of its competence. That trust stemmed not from long term, usual, or positive consequences for control of the system, but rather on immediate, specific, negative consequences for control (Muir and Moray 1996), suggesting the presence of improper trust calibration suggested by Wickens, Gordon, & Liu (1998).

Lee and Moray (1992) found that although performance makes a rapid recovery after faults in the automation, a recovery of trust in the automation was not instantaneous, but took several trials of reliable automation to approach previous levels. More importantly here, the more trustworthy components of an automated system were monitored much less frequently than those that were not trusted.

A study by Kantowitz, Hanowski, and Kantowitz (1997) examined driver acceptance of unreliable traffic information in both familiar and unfamiliar settings. They varied the reliability of their information across three levels: 100% reliable, 76% reliable, and 43% reliable. They found that trust in the information was higher in unfamiliar settings and that the information played a larger part in decisions in unfamiliar settings regardless of the reliability of the information, suggesting the features of a situation that can modulate the mount

of overconfidence. They did not display the calculated reliability of the information in any manner.

Conejo and Wickens (1997) performed a study involving failures of cueing automation on the flight deck. In this study, pilots flew a high fidelity simulator to a specific target that was described verbally and could be located on a map prior to the flight. Once the flight began, the pilot had a 3-D map display that had the target highlighted, the target and a lead-in feature highlighted, or no highlighting. Eight of the conditions had invalid highlighting. Pilots flew to the target area and when they thought they had the target in sight, said "fire", if not in sight, said "abort". Subjects, after locating the target, positioned the target in a reticle and indicated their confidence that they had located the correct target. The trials were scored according to accuracy and confidence in finding the target. They found that scores decreased whenever highlighting was invalid, and that this decrease was due to a loss in accuracy, not confidence. These results show that the pilots retained their confidence/trust in the automation, even while it was failing.

Yeh, Wickens, and Seagull (1998) conducted a similar experiment simulating helmet mounted displays that cued targets of varying importance for army troops. The subjects would scan a virtual environment for tanks, soldiers, mines, and nuclear devices. The soldiers, mines, and tanks were expected events, and the nuclear devices were unexpected, rare events, but had a higher priority for detection than soldiers, mines, and tanks. They found that the presence of cueing aided target detection for expected targets, but with a cost

of drawing attention away from unexpected, but higher priority targets. Again, when the cue was active, it lead subjects to disregard conflicting information in the environment, and rely solely on the automation.

As we have noted, prediction is another type of automation that deals with information integration. The predictor shows the pilot where the automation expects the plane to be in the future. This information must be integrated with information about the plane's current position, and historical information retrieved from memory about the intruder's previous actions. If the pilot maintains trust in that automation at say 99%, then most of the pilot's resources will be put into planning and executing a maneuver to avoid the intruder based upon that predictive display, and little will be put into planning for changes in course or altitude not shown by the prediction. If the predictor is less accurate than the 99%, say about 80%, then using the above strategy, the pilot is over-trusting the predictor and will not be allocating the proper resources to planning for contingencies.

Figure 3 depicts a model of how an operator develops an estimation of automation (e.g. predictor) reliability from past experience with a system. Starting from the bottom of the figure, the operator relies upon long term memory for experience with the system. This memory search can be clouded by recency biases which overweight more recent experiences, and by saliency biases which overweight the most salient experiences while disregarding base rate information (Tversky, and Kahneman, 1974), properties which are consistent with the findings of Lee, Moray, and Muir's work. These experiences are then weighted and combined to form one channel of input into perceived

reliability. The operator would also examine his mental model of the way the system operated in the current situation, and whether those experiences would be valid in this specific circumstance, and use that knowledge as a second channel of input into perceived system reliability. For example, a pilot on the previously mentioned Eastern Airlines flight 401 would have examined his experiences with the autopilot, and finding few recent, salient, experiences where the automation had failed, would have an input from long term memory into his perceived reliability that suggested a high level of reliability. The second input would be from his knowledge of the system, and how it operated, knowing that the malfunction in the light bulb would not have an affect on the reliability of the system. In this manner, his perceived reliability of the system would have been quite high. Once the operator develops a value of perceived system reliability, that reliability influences decision making and selection/planning of possible actions.

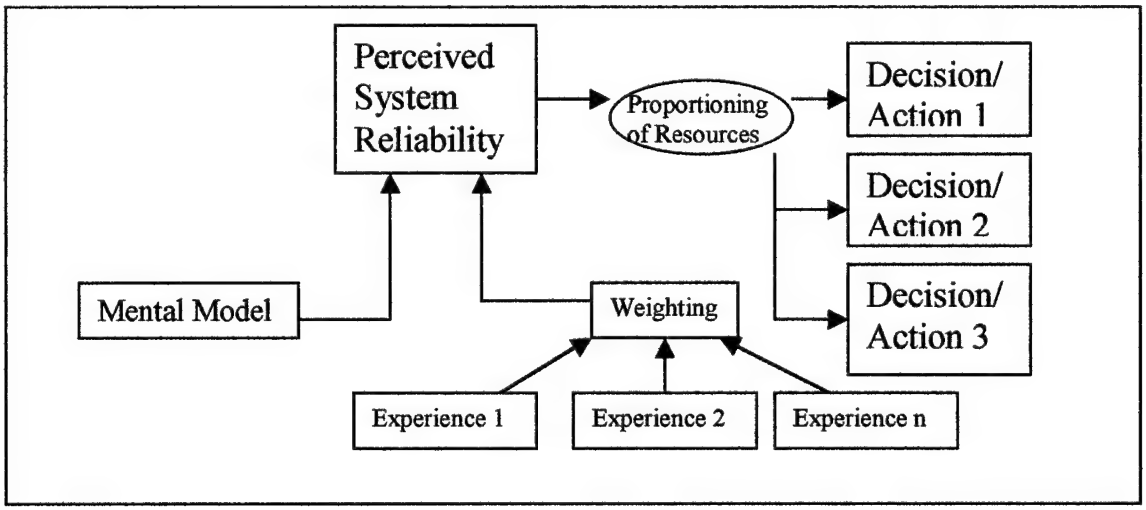


Figure 3: Development of perceived system reliability from previous experience and mental model of system to influence decision making and action selection

Since a multitude of decisions and actions are based on the predictive information contained in the CDTI display, there are serious consequences if

the actions are taken on the basis of incorrect information provided to the pilot by the predictive automation. A maneuver may be made when none is needed, leading to another conflict or simply an inefficiency in the flight. A maneuver may not be made when one is needed, causing a Loss of Separation (LOS). Also, the wrong maneuver may be selected, leading to a loss of separation.

Given that the consequences of improper calibration are known, some effort should be made to facilitate the proper calibration of trust by the operators. Lee and Moray (1994) conclude that designers of any future automated system "should consider how characteristics of the system affect operators' subjective feelings of trust and self-confidence." Specifically, they stress the need "to provide operators with information regarding their performance, and that of the automation, so that operators' self-confidence and trust reflect true capabilities, and promote the appropriate use of automation." This more proper calibration of trust could be achieved through display design, specifically by adding an explicit display of system reliability.

1.5 Display of Reliability

We have identified problems the human operator has in calibrating trust in automation to a level consistent with the automation's reliability. This miscalibration of trust has lead to inefficiencies and safety concerns relating to complacency and over-trust. Now that something is known about human reactions to unreliable automation, an effort needs to be made towards design that affords a proper calibration of trust to actual system reliability. One

method for obtaining this goal would be to make the reliability of the automation explicit on the display. Generally, this type of unreliability estimate has not been used in the previously described research. However, it is seen in both polls (e.g. poll is accurate to +/- 3%) and weather prediction (50% chance of rain).

Surprisingly, only three studies could be located that examined the issue of displayed reliability. Sorkin, Kantowitz, and Kantowitz (1988), used the concept of "likelihood alarms" in a dual task tracking scenario to show that a likelihood information improved attention allocation between the two tasks. The likelihood alarm display, in effect, is a display of reliability, in that it gave information about the certainty of an alarmed system state. The study incorporated two state (alarm/no-alarm) and four state (no-alarm/possible alarm/probable alarm/certain alarm) for a secondary monitoring task, with a primary single axis tracking task. They found that under high workload conditions (difficult tracking), the information about the likelihood of the alarm improved allocation of attention between the two tasks, without increasing the operator's attentional load over the two-state alarm.

Montgomery, and Sorkin (1996) used a visual display of reliability to help subjects weight the importance of each of 9 graphical elements in a display. The subjects' task was to decide whether the information displayed by the 9 elements of the display showed a signal or noise. Each element of the display was an independent sample from either a signal or noise distribution, and the elements' reliability was varied by changing the variance of the

distribution. They found that subjects tended to weight high reliability sources slightly more than low reliability sources, and their detection performance and weighting efficiency were the best with a reliability display that varied the luminance of the particular display element, with brighter elements being more reliable than dimmer elements.

Another example of the use of graphical representations of uncertainty/reliability was examined by Kirschenbaum and Arruda (1994). In their study, an algorithm was developed to display a 95% uncertainty ellipse that described the area of a target position to submarine operators. The algorithm used the same rule set that uses the same information as current quantitative verbal indicators. In target acquisition, the uncertainty comes from the reliability of information given to the display by sonar. Sonar cues are then used by the operators to determine the range, speed, and course of the target. The size of the ellipse starts out large, and decreases in size as more information is gathered and the accuracy of radar data is increased with decreasing target range. Ideally, the operator could delay firing on the target, or avoiding the potential obstacle until the ellipse becomes a point, however, in this scenario a delay of action could lead to a collision with the target or allow the target to fire its weapons first. Therefore, allowing the operator to make a decision to fire or maneuver as early as possible was a goal of the designers in developing the ellipse display. Subjects were highly experienced instructors for the display type the authors used in the study. Range error, scenario time, and confidence of the target's range were the primary dependent variables.

Range error was the absolute difference in yards between the operator's range estimate of the target at the time they fired their weapons and the actual range at the same time. Confidence in their range estimate was measured on a 5 point Lickert scale. Scenario time initiated at the beginning of the scenario and ended at the command to fire weapons. In the difficult scenarios, subjects were significantly more accurate in giving range estimates at the time of firing weapons with the graphical display, than those using the verbal uncertainty display, while there was no difference in the easy scenarios. There was no significant difference in confidence or scenario time. The authors concluded that the graphic (spatial) representation of uncertainty can contribute to decision performance in a spatial problem. We propose to examine this concept of an explicate graphical display of reliability in the prediction function of a CDTI which is used in the spatial problem of traffic avoidance and navigation.

Figure 4 shows the effects of a display of actual system reliability could be incorporated into the model of perceived system reliability suggested in figure 3. Starting at the top of the figure, information channels which are used by the automation to develop a predictor are each shown with their associated reliability. These individual reliability values are then combined to form an overall system reliability measure which is displayed explicitly to the operator. This displayed system reliability is then combined with the products of knowledge and experience (from the earlier model shown in figure 3) to form the operator's perceived system reliability which is used to make decisions and actions. In this case, as compared to the model presented in figure 3, the

human should have a more accurate representation of system reliability, therefore being able to more properly calibrate his trust in the system to its actual reliability, leading to better decisions, more appropriate actions, and better preparation for alternatives.

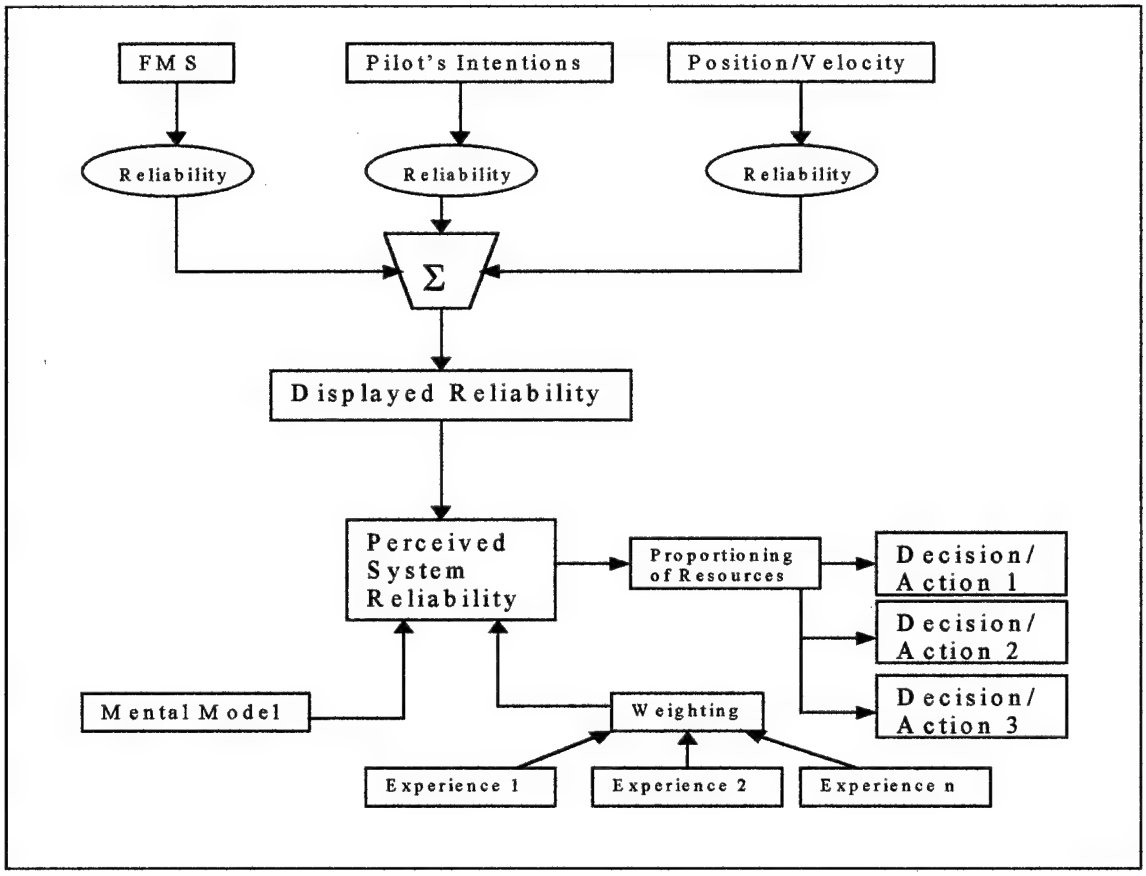


Figure 4: Addition of explicit display of system reliability to formation of operator perceived system reliability

There are, however, two types of automated prediction failures that can create an invalid predictor. A valid predictor is one in which the intruder follows the predicted path perfectly. An invalid predictor can have the intruder depart from the predicted path in two ways. The intruder can depart the predicted path within or outside a certain displayed error range. Figure 5

shows how a predictor for an intruder aircraft with a graphical uncertainty wedge would look. It shows the aircraft's current location, its best estimate of future location, as well as the limits of 95% confidence interval of that future location. An aircraft that deviated within those limits (about 95% of the time), would have the path indicated by the actual future position line.

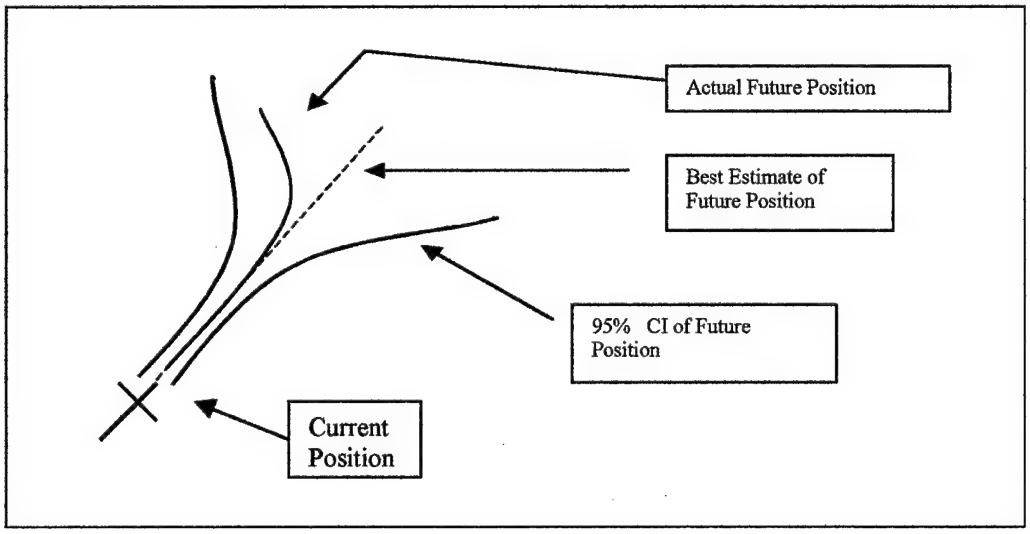


Figure 5: Deviation within displayed reliability (Adapted from figure 6.7 Wickens, Mavor, Parasuraman, & McGee 1998)

Figure 6 shows the same display, however in this case, actual future position of the aircraft falls outside of the 95% confidence interval, an event that would occur about 5% of the time.

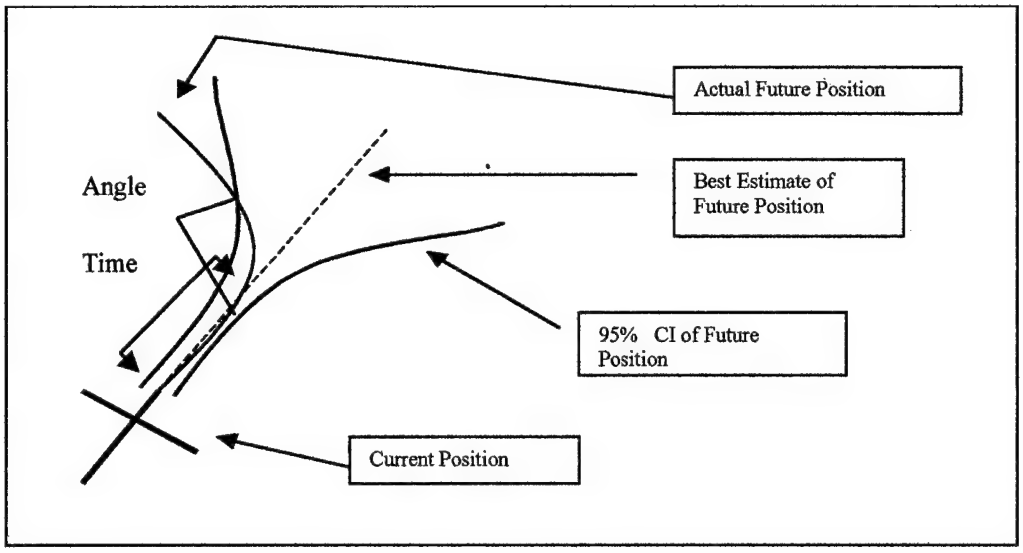


Figure 6: Future Position outside of displayed reliability (Adapted from figure 6.7 Wickens, Mavor, Parasuraman, & McGee 1998)

In either case, the intruder could depart from the predicted path in a manner that either decreases the time to loss of separation (LOS) or increases the time to loss of separation (possibly eliminating that loss entirely). Variables that influence the departure from predicted course include time to departure and angle of departure. Displaying the 95% confidence interval of where the intruder will be in the future, presents the integrated reliability of the automation to the pilot in graphical position terms that can be used as a basis for making maneuver decisions. This estimate of reliability, which is critical for determining the width and shape of the 'wedge' in figures 4 and 5, could be calculated using a number of methods. For the purpose of this study, we assume that it would be based on some type of random sampling of a large number (say 10,000) aircraft tracked on a radar display with known present position and velocity as well as FMS information, etc. that would be available for use in generating a predictor. Then these aircraft are sampled at several

points in the future, and a 95% confidence interval of their future position would be calculated and used to generate the reliability estimate for the predictor and hence the "splay angle" of the wedge as shown in figure 7.

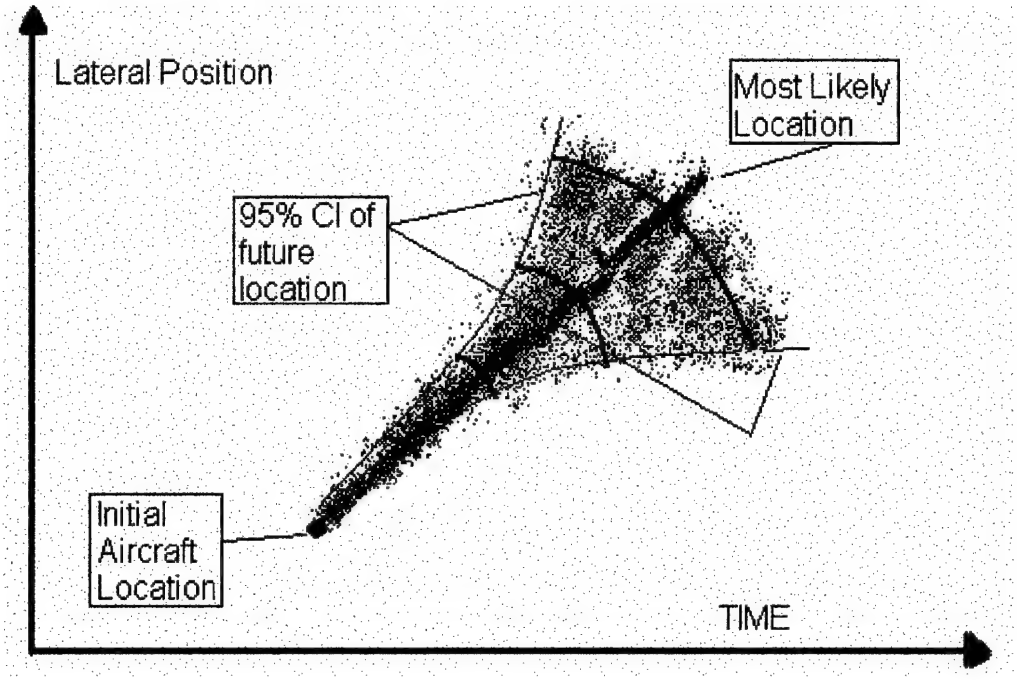


Figure 7: Concept of reliability estimate generation, where each point represents the future position of an individual aircraft plotted relative to an initial location.

1.6 Integration into Current Study

Previous studies have investigated many of the pertinent issues to the display of traffic on the CDTI. The general benefits of predictive displays on tracking and specifically on the CDTI in the free-flight scenario have been supported (Wickens and Morphew, 1997; Jensen, 1981; Jago and Palmer, 1982; Johnson, et. al., 1997; Barhydt & Hansman, 1997; Merwin & Wickens, 1996). However, no CDTI studies have included a less than perfectly reliable predictor, although it is quite certain that this automation feature *can not be*

perfectly reliable given the sources of information used to generate the predicted path of the aircraft. Unreliable automation impacts trust, which in turn has implications for human performance and use of automation tools (Parasuraman and Riley, 1997; Muir 1987; Lee and Moray 1992, 1994; Muir and Moray, 1996; Kantowitz et al, 1997). Although individual differences have been implicated in automation use under less than perfectly reliable conditions, the design of the automation itself could have an impact across individuals on how automation, such as a predictive CDTI, is used in the cockpit.

Specifically, the operator must develop some sort of automation use strategy, and preparation for alternatives, based upon a subjective assessment of the reliability of the automation. That subjective assessment can come from previous experience with the automation and internal knowledge representations of both the external situation and the automation itself as in figure 3. By graphically displaying the automation reliability to the operator, he can compare his knowledge and experience with the display to form a better assessment of the true reliability of the system as in figure 4. In turn, that more accurate perceived system reliability should be put to use in decision making, preparation, and action planning, making the operator less prone to complacency. However, few studies have examined the effects of a reliability display on performance, and none have examined it in the concept of the CDTI.

Using a coplanar CDTI with previously proven display symbology, the current study will attempt to add further understanding to the benefits of predictive information on CDTI previously found by Barhydt and Hansman

(1997), Wickens and Morphew (1997), and Johnson et al (1997), by addressing issues concerning the display of the reliability of that predictive information and the user reaction to that displayed reliability in a less than perfectly reliable system.

In the present study, participants will perform a simulated flying task, using a CDTI with predictive information to navigate to a 3D waypoint in space, while avoiding potential conflicts with an intruder aircraft, and performing a secondary monitoring task. Throughout the simulation, a perfectly reliable predictor will be shown on their own aircraft and a predictor that is reliable on 83% of the trials will be shown on the intruder aircraft. This unreliability would represent a case, for example, when ownship's CDTI predictor symbol was based upon FMS information from the intruder aircraft, but the pilot of the intruder aircraft decided to deviate from the programmed FMS path. On one trial out of six, this change from a perfectly reliable predictor to a less than perfectly reliable predictor will be the significant difference from prior work on CDTI displays. Two types of displays will be used as a between subjects variable. The first display type will show only the single line (SL) predictor, while the second will incorporate a graphical representation of the reliability of the predictor, in much the same manner as that used by Kirschenbaum and Arruda (1994). Instead of an ellipsoid as a graphical representation, we assume that the exact location of the intruder is known, but the future location is estimated with decreasing certainty as the span of prediction grows, so a 'wedge' shaped (W) predictor (Figure 5), indicating a 95% confidence interval of future locations, both horizontal and vertical, will be used as the second display type. A variety of approach geometries by the intruder will be presented throughout the simulation,

including level, descending, and ascending approaches from 45, 90, and 135 degrees both right and left of ownship.

In the SL display, the operator will be using only the internal representation of predictor reliability (model from figure 3), while in the W display the operator will have the external reliability information as well (model from figure 4). Since the invalid predictor would be a more expected event in the W display case, pilots should put more processing resources into planning for that failure, and should have better performance responding to the intruder's course change from the predicted course than pilots using the SL display. Pilots using the W display would plan for some changes in intruder trajectory that depart from the predicted course, making maneuvering decisions based on that knowledge and performing better on trials where the predictor is invalid compared to pilots using the SL display.

In a trial with an invalid predictor, the secondary monitoring task performance should suffer, because more resources will need to be diverted from monitoring to making maneuver decisions and inputs to the aircraft controls. In the W display type, that drop in monitoring performance should not be as great, because the invalid predictor would be more of an expected event, and if planned for, require fewer resources than the unexpected event in the S display case. It is possible, however, that such greater planning for alternatives during the reliable trials might impose a greater resource cost for the wedge display, reflected in a loss of secondary task performance on those trials.

2.0 Experimental Method

2.1 Participants

The 20 participants were all licensed pilots with instrument ratings, most from the University of Illinois Institute of Aviation. They received $6 per hour for their participation. The following table shows descriptive statistics of ages, number of flight hours, and number of instrument hours.

Variable	Mean	Std Dev	Minimum	Maximum
AGE	26.65	10.01	19.00	54.00
HOURS	1639.85	3689.28	130.00	12500.00
INSTRUMENT	205.50	444.92	20.00	2000.00

In addition, 17 of the 20 were instrument current, and half of the participants were certified flight instructors. There were 5 females and 15 males.

2.2 Simulation Flight Dynamics and Apparatus

The simulation ran on a Silicon Graphics workstation and was viewed on a 20-inch Silicon Graphics color monitor with a screen resolution of 1280 X 1024 pixels at 25 frames per second. The pilots controlled the simulation using a two-button joystick which varied pitch, bank, and airspeed. Limits of +/- 5 degrees of pitch and +/- 30 degrees of bank limited maneuvers to those normally used in passenger flight. Constant rate speed changes were controlled by the top button (faster) and trigger (slower). A linear coupling function was also included in the flight simulation dynamics, causing cross coupled responses to input from the flight controls (e.g. pitching up or down resulted in a decrease or increase in airspeed, and banking resulted in a pitch down) as is found in real flight. The target speed was 325 kts, with a maximum speed change capability of +/- 150 kts. Initial conditions in each scenario had the ownship heading 360

degrees at 10,000 ft and 325 kts. Pilots were instructed to fly to a 3D waypoint, and to deviate from their original altitude, heading, and airspeed as little as possible while avoiding conflicts and proceeding to the next waypoint. Light turbulence was programmed into the simulation, causing the aircraft to drift slowly from the prescribed heading and pitch angles, thus forcing the pilots to maintain active control of the aircraft at all times.

2.3 Task and Simulation

Pilots flew four sets of 15, 1-2 minute flight scenarios, with a short break between each set. In each scenario, pilots were required to fly to a designated navigational waypoint (depicted as a VOR symbol on the display) directly ahead of the ownship without coming into conflict with an "intruder" aircraft located in their airspace, and minimizing deviations from prescribed heading, altitude, and airspeed. A conflict is described as an intruder aircraft entering the protected zone of the ownship, which was a cylinder of airspace with a radius of 3 miles and an altitude of +/- 1500 ft. If an intruder entered that airspace, this would be considered in conflict. Pilots were instructed to avoid conflicts. The task required the pilots to determine whether the intruder's flight path would penetrate their protected zone and then maneuver to avoid the conflict. Each scenario contained only one intruder aircraft.

Each scenario was initiated as a conflict or non-conflict trial (C or NC). A conflict trial would put the two aircraft into a conflict if neither changed course or airspeed from the start of the trial. A non-conflict trial would result in no conflict if neither aircraft changed course or airspeed from the start of the trial. There were 50

conflict trials and 10 non-conflict trials. To simulate a less than perfectly reliable predictor, 10 of the 60 trials had the intruder aircraft change heading or rate of climb/ descent. Heading changes were between 16.5 degrees and 33.3 degrees, and changes in pitch were from a climb or descent to level. Changes in trajectory occurred at unpredictable times between 14 and 38 seconds (mean = 24.3 seconds) after the trial started. Changes in trajectory were not reflected by the automation generating the prediction. That is, the predictor line would continue to point in the direction of the original flight path. Of the 10 change trials (Δ), 6 were initially conflict (C) trials and 4 were initially non-conflict (NC) trials. The changes would either make the conflict easier to avoid (L-less maneuvering required), or more difficult (M-more maneuvering required). This design can be seen in figure 8. The greatest focus of the study will concentrate on the 8 trials in which more maneuvering is unexpectedly required.

	NΔ	L Δ M	Total
C	44	1 6 5	50
NC	6	1 4 3	10
Total	50	2 10 8	60

Figure 8: Frequency of trial type

2.4 Displays

The design of the display was much like that used in previous CDTI experiments, specifically Wickens and Morphew (1997). It was a 2-D coplanar display with a top-

down and forward looking view of the surrounding airspace. The display included an attitude indicator in the top center, which allowed monitoring of pitch and bank attitudes. Two vertical tapes displaying the altimeter (right) and the airspeed (left) were positioned on either side of the coplanar CDTI. Along the top of the display, a strip for forward field of view (FFOV) simulated an 'out of the cockpit' scan area where small ellipses (FFOV indicators) appeared at random times and in random locations along the strip to simulate visual scanning demands. The FFOV indicators appeared for a 15 second period, or until noticed and acknowledged by the pilot with a spacebar press. Three or four indicators appeared during each approximately 90 second trial. The FFOV indicators were configured so as not to be detectable by peripheral vision, but had to be directly observed to be detected, thereby forcing a head-up scanning pattern. The pilots were to respond to the appearance of an ellipse by hitting the spacebar on a keyboard in front of them, at which time the ellipse disappeared.

The CDTI consisted of a horizontal situation indicator (HSI) which portrayed a top down (x-z axes) view of the pilot's surrounding airspace. Below the HSI was a forward-looking (x-y axes) view of the same airspace. The traffic symbology was overlaid on a grid of equi-spaced lines representing 5 nautical mile increments. The lines were composed of white dots spaced at intervals of 1 nautical mile. The grid rotated with the ownship to provide consistent spacing information of traffic symbology. The traffic symbology in the HSI contained aircraft symbols consisting of ownship (magenta) and the intruder aircraft (gray), and a navigational waypoint consisting of a blue VOR map symbol. Two types of displays were used. The first had a single prediction line coming off of each aircraft that depicted the aircraft's location 45 seconds into the future (SL-

Single Line Predictor) . The predictor lines were the same color as the aircraft symbols. The second condition included the same predictor lines on both aircraft, but added two curved lines to the intruders predictor, one on either side of the predictor line in the shape of a wedge, that indicated an interval of possible future locations along the predicted path. (W- Wedge Predictor). The lines represented a 95% confidence interval of the aircraft's future location and the width of that interval increased as the prediction span increased, indicating to the pilot that 95% of the time the pilot could expect the intruder to be between these boundaries in the future. The change (Δ) trial parameters were set so that this estimate would be approximately correct. The lines were generated using a parabolic function to give them their shape. The two display conditions can be seen in figure 9.

a)

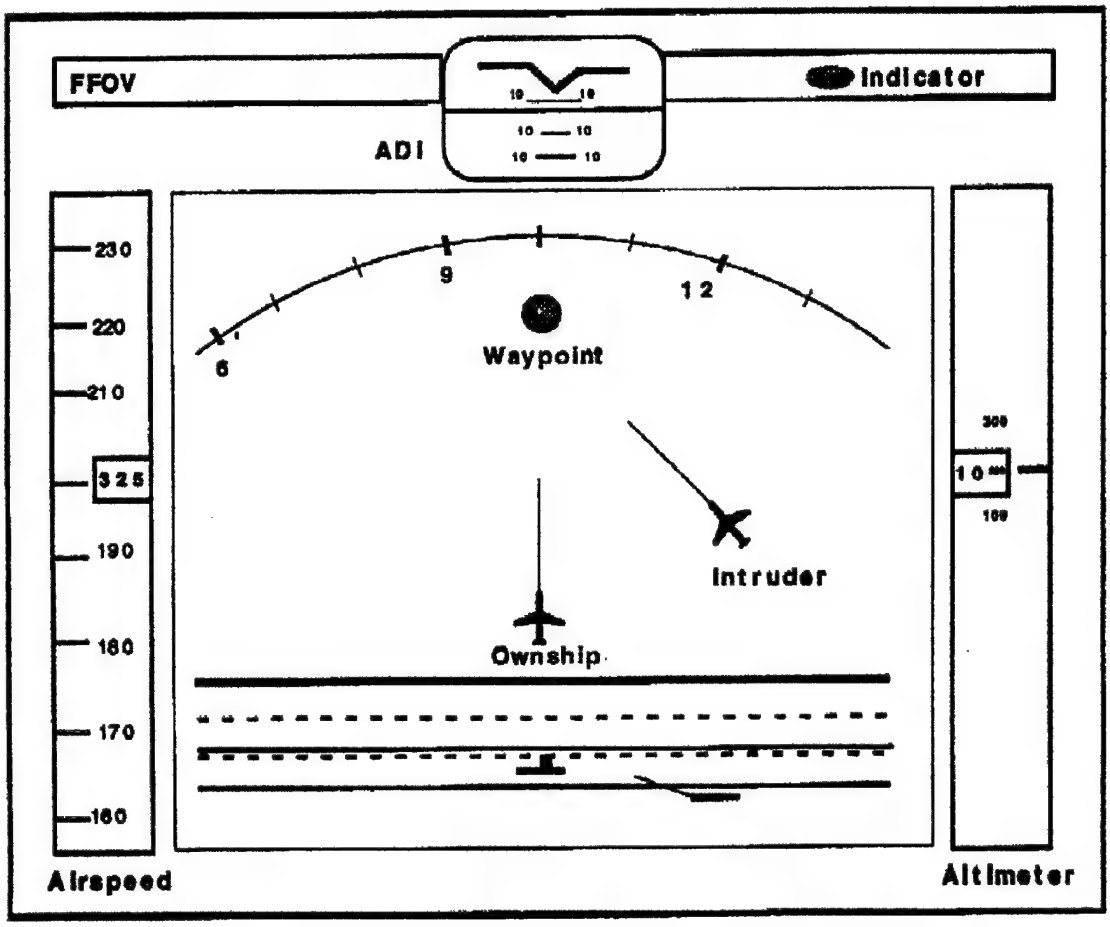


Figure 9a: Single line (SL) predictor display

b)

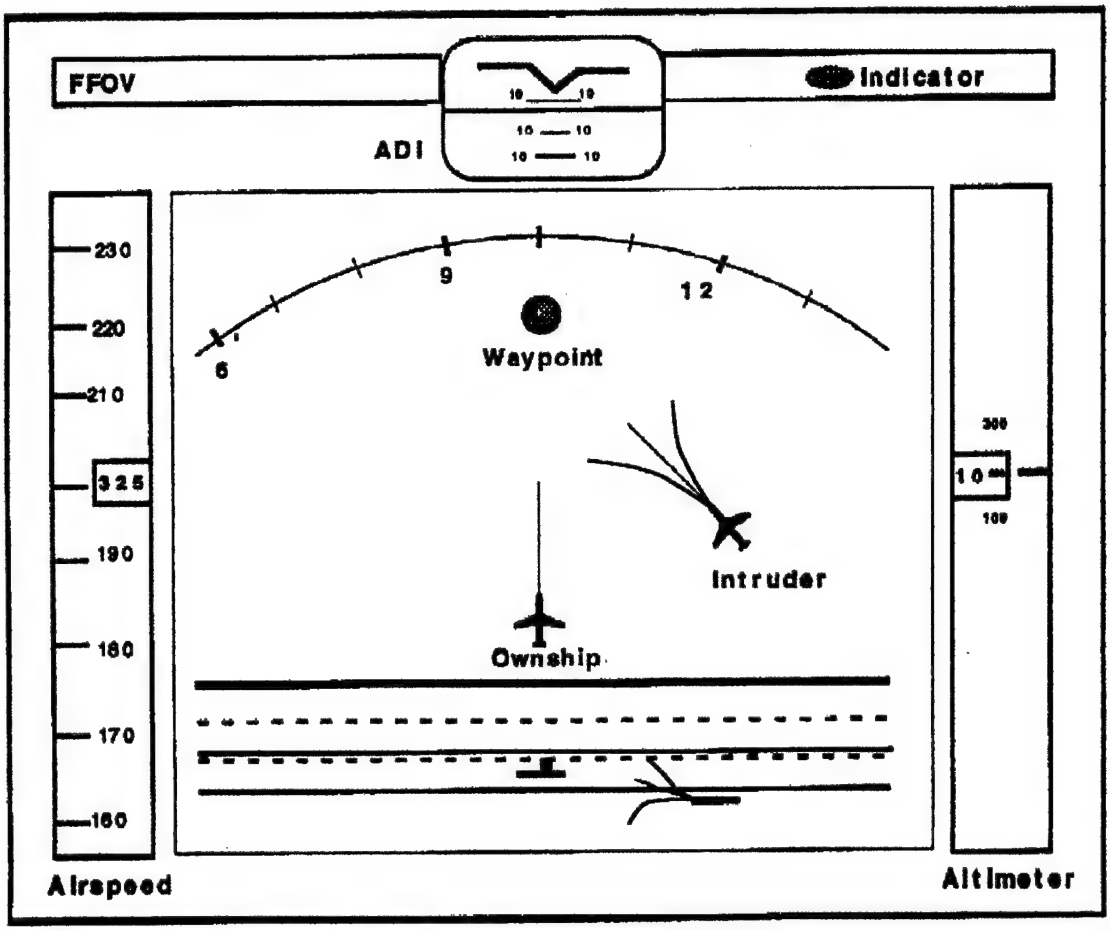
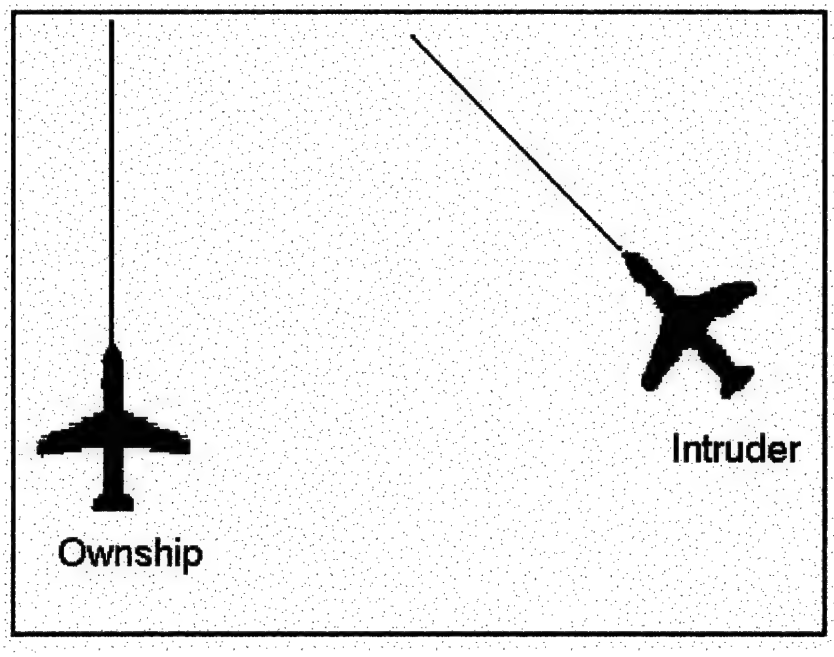


Figure 9b: Wedge predictor (W) display

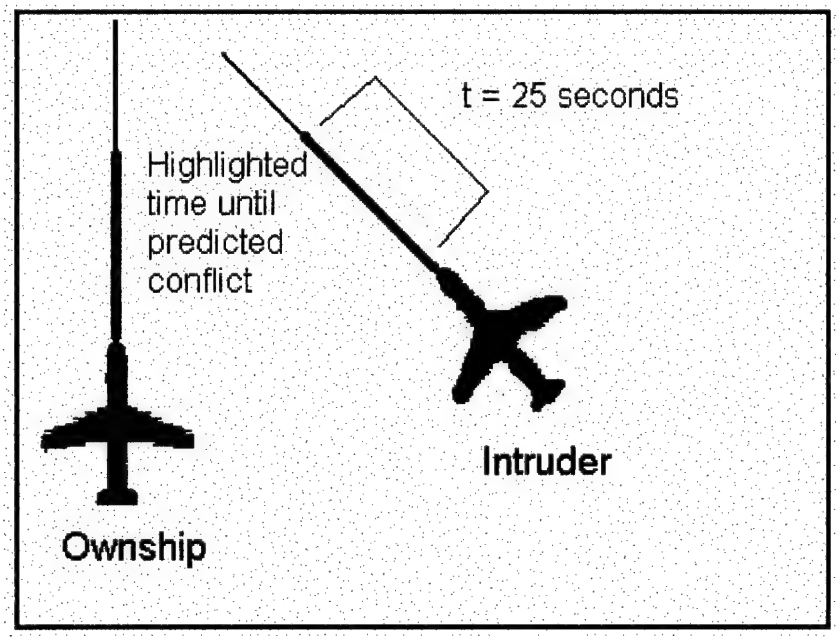
In addition, the predictor lines on both aircraft were highlighted in the case of a predicted conflict. The length of the highlighted portion of the predictor lines decreased in the direction of the aircraft as the time to predicted conflict decreased, thus giving the pilot a direct estimation of the time to loss of separation. Once a predicted conflict became an actual conflict, the intruder would light up yellow. Figure 10 illustrates the

display for no conflict, a predicted conflict, and an actual conflict in the single line (SL) display case.

a)



b)



c)

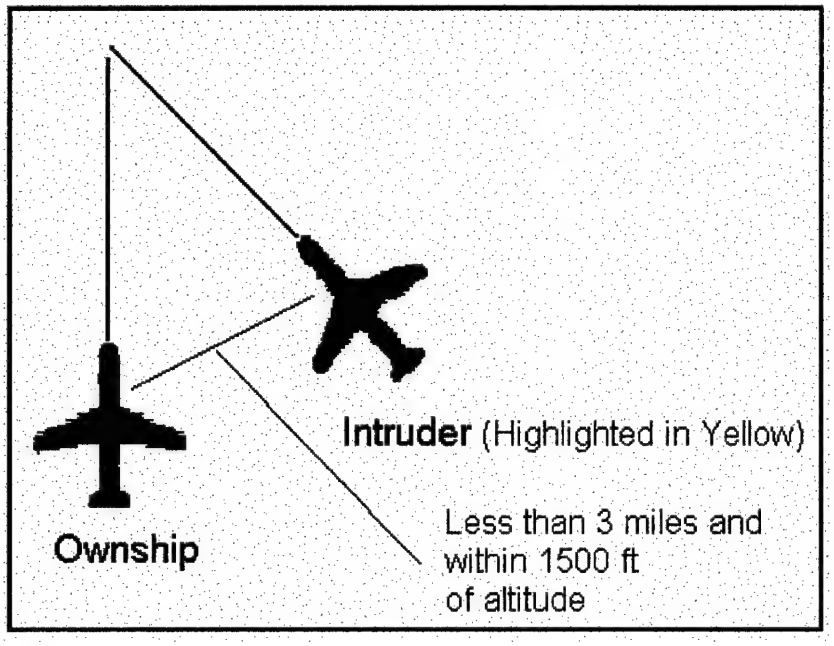


Figure 10: a) no conflict, b) predicted conflict, c) actual conflict

The forward (profile) view of the coplanar display contained a set of parallel horizontal yellow lines representing the ownship's current vertical protection zone boundaries (1500 ft above and below the aircraft). Dashed yellow lines indicated the location of the protected zone boundaries 45 seconds into the future. Each aircraft also had its associated predictor lines and wedge (in the W display type) in the vertical plane as well. The forward view is depicted at the bottom of figures 8 and 9.

For both the top down and forward views in Δ trials, when the intruder changed its trajectory contrary to that of the predictor, the predictor remained at its original (pre-change) bearing and rate of climb/descent from the aircraft. This simulated a predictor based on some combination of both present position/velocity and information from the intruder's flight management system (FMS). The actual trajectory then changed from the

predicted trajectory, simulating the intruder aircraft changing heading or rate of climb/descent without updating FMS information, at which time the predictor would not be valid for the intruder's actual trajectory. Two examples are shown in figure 11. In the altitude change part of the figure, the aircraft's predictor is indicating a descent, while the aircraft remains at a level altitude. In the lateral position change, the aircraft's predictor indicates a direct path, but in actuality turns left while the predictor remains at a constant heading.

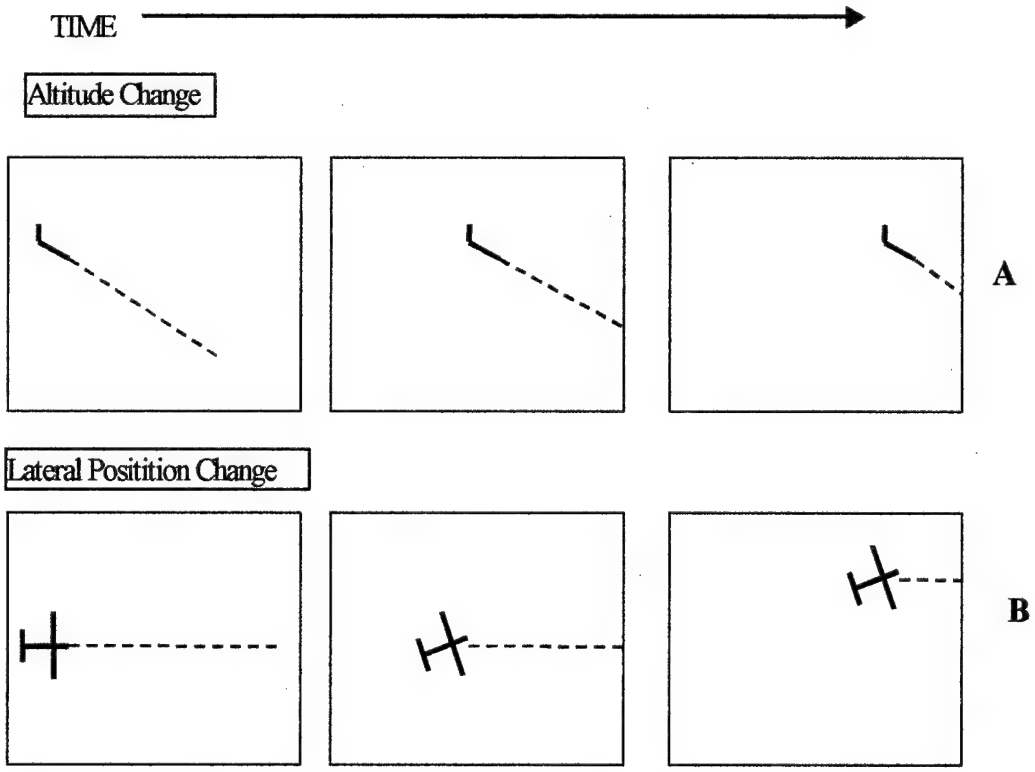


Figure 11: Two examples of three time frames showing an unreliable predictor

2.5 Experimental Design

A 2x3x3x2 factorial mixed design was used. Display type (SL or W) was varied between subjects, while vertical traffic geometry (ascending, level, descending) longitudinal geometry (45, 90, and 135 degrees), and trial predictor validity, were varied

within subjects. 10 trials having an unpredicted trajectory change by the intruder, were distributed within the set of 60 trials, thus rendering the predictor invalid for that trial and achieving an overall reliability of 83.3 percent for the entire series of 60 trials. The trajectory change trials were distributed in such a manner as to appear random to the pilot. Pilots were assigned to each display group in a manner to ensure both groups had roughly equal numbers of more and less experienced pilots. Both groups of subjects saw the exact same sequence of trials including 50 predictor valid trials and 10 predictor invalid trials.

The intruder aircraft approach geometries were varied in order to ensure exposure to a variety of traffic patterns, including three vertical and three horizontal geometries, with approaches from both right and left of ownship.

2.6 Procedure

Subjects participated in one session consisting of four sets of 15 trials each. Before the session, subjects were read instructions and shown illustrations of the task and corresponding displays. They were specifically told that the predictor display was not 100% accurate. After instructions, the participants flew 10 practice trials with a perfectly reliable predictor, to become comfortable with the displays. The experimenter sat in the simulation room with them during the practice session to answer any questions online. Upon completion of the practice, subjects were invited to ask any further questions and then began the first set of trials. Between each set of trials, subjects were required to take a short (3-5 minute) break before continuing with the simulation to avoid any fatigue effects.

Upon completion of the final set of trials, pilots were given a questionnaire which asked for their subjective estimate of the reliability of the predictor, and their preferred avoidance maneuvers and strategies. Finally, subjects were asked for any additional comments, thanked for their participation and paid for their time.

3.0 Results

Prior to any analysis, data were examined for gross deviations from normality and outliers. No data points appeared to be outliers, and any transformations used will be annotated in the analysis. All statistical tests were performed using SPSS student version 6.1 for Windows.

3.1 Primary Task Measures: Traffic avoidance

Primary task measures included three types. The first type consisted of safety measures which included the total time pilots spent in predicted conflicts and actual conflicts during each flight. The second type consisted of efficiency measures, which examined RMS deviations from 0 degrees heading, 325 knots airspeed, and 10,000 ft of altitude. The final type consisted of maneuver response measures, which were the time between the start of the trial and when the first maneuver was made, and the type (airspeed change, heading change, or altitude change) of the first maneuver made. In this case the first maneuver was defined as an airspeed deviation of 30 kts, and altitude deviation of 500 ft or a heading change of 10 degrees. These criteria were derived in the following manner. Generally, when pilots maneuvered to avoid a conflict, they either changed heading by 30 degrees or more, changed airspeed by 90 kts or more, or changed altitude by 1500 ft or more. The maneuver criteria were then set at about 1/3 the size of the maneuver deviations to ensure that the pilot had committed to that type of avoidance maneuver before it was recorded as such.

3.1.1 Total Time in Predicted and Actual Conflicts

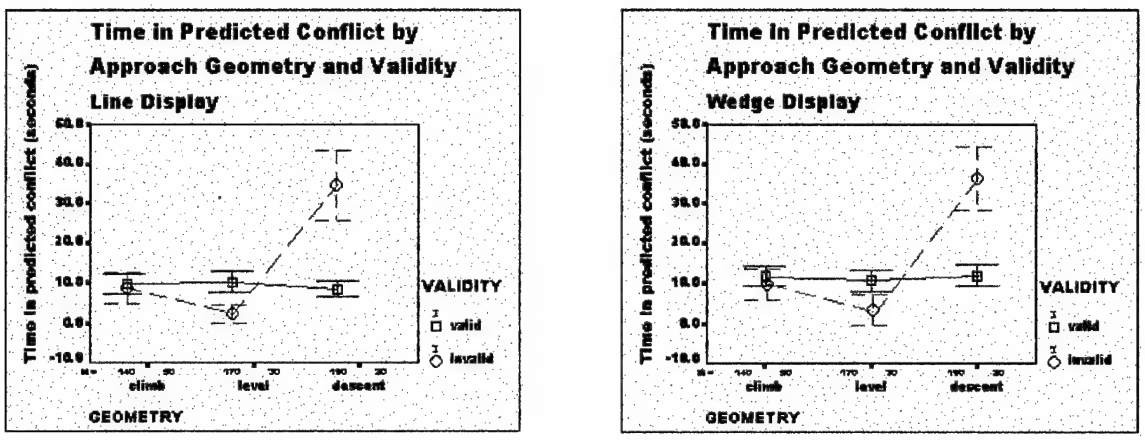


Figure 12: Time in predicted conflict measures (bar represents 95% confidence interval) Left panel is line display and right panel is wedge display

Time in predicted conflict

by VALIDITY DISPLAY GEOMETRY

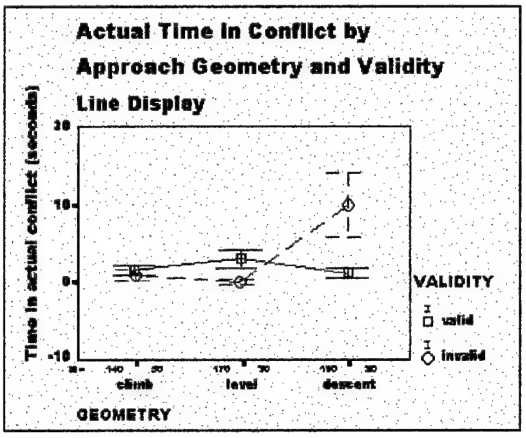
UNIQUE sums of squares

All effects entered simultaneously

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig of F
Main Effects	246300163	4	61575040.826	25.847	.00
VALIDITY	39951262	1	39951261.982	16.770	.000
DISPLAY	5097539	1	5097538.849	2.140	.144
GEOMETRY	232521786	2	116260893.149	48.802	.00
2-Way Interactions	244249929	5	48849985.730	20.505	.00
VALIDITY DISPLAY	61653	1	61652.516	.026	.872
VALIDITY GEOMETRY	242790243	2	121395121.390	50.957	.00
DISPLAY GEOMETRY	1234635	2	617317.454	.259	.772
3-Way Interactions	310595	2	155297.675	.065	.937
VALIDITY DISPLAY GEOMETRY	310595	2	155297.675	.065	.937
Explained	292466653	11	26587877.542	11.161	.00
Residual	2830182222	1188	2382308.267		
Total	3122648875	1199	2604377.710		

Table 1: ANOVA table of time in predicted conflict

Figure 12 depicts time in predicted conflict per trial, where there was one main effect of approach geometry [$F(2,663)=11.005$, $p<.001$] and an interaction effect between predictor validity and approach geometry [$F(2,663)=20.343$, $p<.001$] (see table 1). The main effect of geometry really only reflects the large effect found in the interaction with predictor validity as seen in the right-most points of both panels of figure 12. Here, the cost of invalid trials was only reflected in trials with a descending approach geometry. Invalid/descending trials had a mean time in predicted conflict of 35.06 seconds per trial, while valid/descending trials had a mean time in predicted conflict of 10.31 seconds per trial.



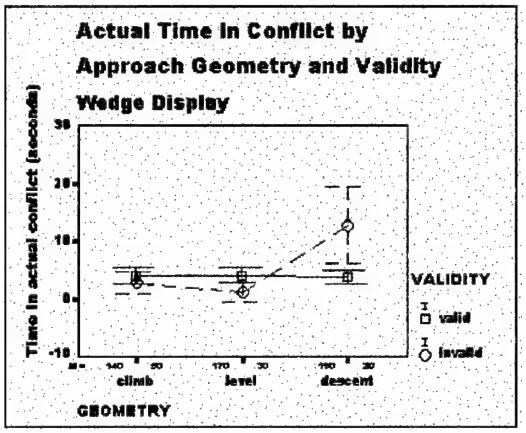


Figure 13: Time in actual conflict measures (bar represents 95% confidence interval)

* * * A N A L Y S I S O F V A R I A N C E * * *

Time in Actual Conflict by VALIDITY DISPLAY GEOMETRY

UNIQUE sums of squares

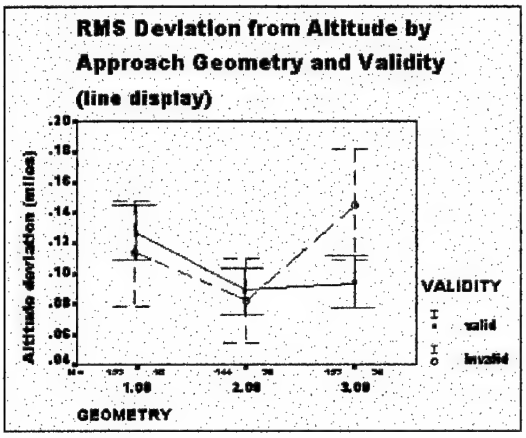
All effects entered simultaneously

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig of F
Main Effects	30604894	4	7651223.591	14.657	.00
VALIDITY	4222461	1	4222461.398	8.089	.005
DISPLAY	5186548	1	5186548.096	9.936	.002
GEOMETRY	24850327	2	12425163.633	23.802	.000
2-Way Interactions	33088239	5	6617647.704	12.677	.00
VALIDITY DISPLAY	300	1	299.842	.001	.981
VALIDITY GEOMETRY	32501564	2	16250781.830	31.131	.00
DISPLAY GEOMETRY	582072	2	291035.978	.558	.573
3-Way Interactions	39673	2	19836.595	.038	.963
VALIDITY DISPLAY GEOMETRY	39673	2	19836.595	.038	.963
Explained	46419167	11	4219924.292	8.084	.00
Residual	620155958	1188	522016.799		
Total	666575125	1199	555942.556		

Table 2: ANOVA table of time in actual conflict

Figure 13 depicts the measure of pilots' time in actual conflict per trial, which showed a main effect of all three variables. There was a main effect of validity [$F(1,1188)=8.089$, $p=.005$], display [$F(1,1188)=9.936$, $p=.002$], and of geometry [$F(2,1188)=23.802$, $p<.001$] , as seen in table 2 which depicts the ANOVA table for time in actual conflict. The wedge display lead to more time in conflict than the line display (mean= 4.1 and 2.2 seconds respectively). However, the effects of geometry and validity can, once again, be best described in terms of their interaction [$F(2,1188)=31.131$, $p<.01$], which can be seen on the right of both panels in figure 13. In this case, as in the predicted conflict case, the cost of the invalid predictor (more time in conflict) appeared only on descending trials. The cost of invalidity on descending trials was 8.85 more seconds in conflict per trial (mean valid/descending=2.63 seconds, mean invalid/descending=11.48 seconds per trial). This means that the pilots spent 23% more time in an actual conflict with the intruder aircraft on descending/invalid trials than descending/valid trials, showing the particular difficulty of this combination of independent variables.

3.1.2 Efficiency/Deviation measures



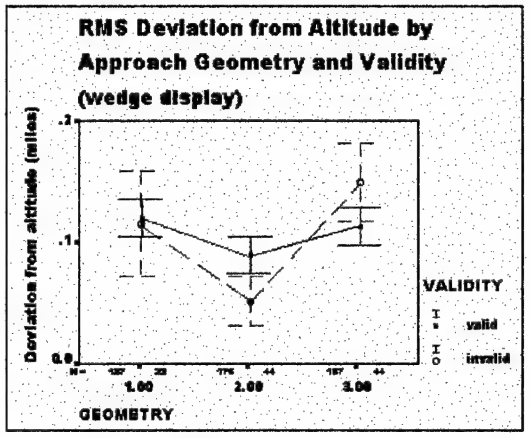


Figure 14: RMS Deviation of altitude measure (bar represents 95% confidence interval)

* * * A N A L Y S I S O F V A R I A N C E * * *

Log(RMS deviation from Altitude)

by VALIDITY

DISPLAY

GEOMETRY

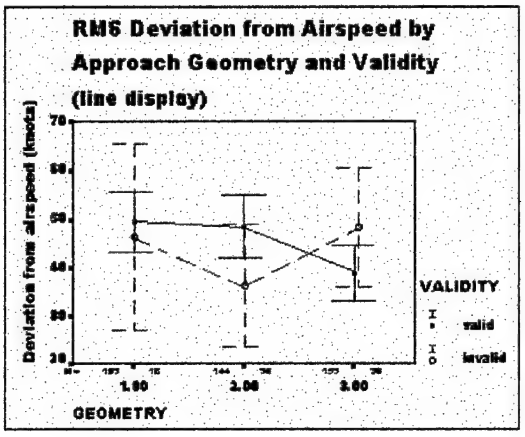
UNIQUE sums of squares

All effects entered simultaneously

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig of F
Main Effects	12.350	4	3.088	7.217	.000
VALIDITY	2.675	1	2.675	6.252	.013
DISPLAY	.100	1	.100	.235	.628
GEOMETRY	10.684	2	5.342	12.487	.000
2-Way Interactions	6.723	5	1.345	3.143	.008
VALIDITY DISPLAY	.347	1	.347	.811	.368
VALIDITY GEOMETRY	4.970	2	2.485	5.809	.003
DISPLAY GEOMETRY	1.173	2	.586	1.371	.254
3-Way Interactions	.186	2	.093	.217	.805
VALIDITY DISPLAY GEOMETRY	.186	2	.093	.217	.805
Explained	18.968	11	1.724	4.031	.000
Residual	508.235	1188	.428		
Total	527.203	1199	.		

Table 3: ANOVA table of Log (RMS deviation from altitude)

Figure 14 depicts the RMS deviation from the prescribed altitude of 10,000 ft in a scale of miles per trial. RMS deviation from altitude was analyzed using a log transformation to normalize the distribution. As shown in table 3, which shows the ANOVA table of Log(RMS deviation from altitude), there were two main effects, that of predictor validity [$F(1,1188)=6.252$, $p=.013$] and of approach geometry [$F(2,1188)=12.487$, $p<.001$], as well as an interaction between predictor validity and geometry [$F(2,1188)=5.809$, $p=.003$]. Climbing, and descending trials (mean=634 ft per trial and 581 ft per trial respectively) had a larger deviation than level trials (mean=422 ft per trial). The interaction between validity and geometry, seen in figure 14, shows that invalid trials produced greater deviations from altitude than valid trials when the intruder was descending, a pattern similar to that observed with the safety measures of time in predicted and actual conflict. Invalid/descending trials had a mean deviation of 792 ft per trial, compared to valid/descending trials which had a mean deviation of 528 ft per trial.



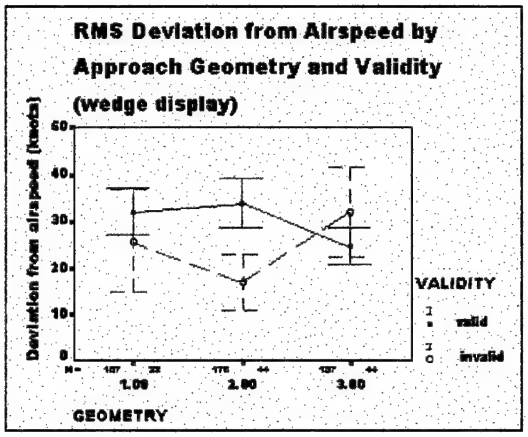


Figure 15: RMS deviation of airspeed measure (bar represents 95% confidence interval)

* * * A N A L Y S I S O F V A R I A N C E * * *

Log(RMS deviation from Airspeed)

by VALIDITY

DISPLAY

GEOMETRY

UNIQUE sums of squares

All effects entered simultaneously

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig of F
Main Effects	10.230	4	2.557	7.752	.000
VALIDITY	.074	1	.074	.225	.636
DISPLAY	7.346	1	7.346	22.264	.000
GEOMETRY	2.443	2	1.222	3.703	.025
2-Way Interactions	6.515	5	1.303	3.950	.001
VALIDITY DISPLAY	.299	1	.299	.907	.341
VALIDITY GEOMETRY	6.060	2	3.030	9.183	.000
DISPLAY GEOMETRY	.119	2	.059	.180	.835
3-Way Interactions	.060	2	.030	.091	.913
VALIDITY DISPLAY GEOMETRY	.060	2	.030	.091	.913
Explained	19.689	11	1.790	5.425	.000
Residual	391.945	1188	.330		
Total	411.634	1199	.		

Table 4: ANOVA table of RMS deviation from airspeed

RMS deviation from airspeed was analyzed using a logarithmic transformation to make the distribution appear more normal. As seen in table 4, which is the ANOVA table of Log(RMS deviation from airspeed), there was a main effect of display type [$F(1,1188)=22.264$, $p<.001$], with the wedge display having less deviation (mean=29.3 kts per trial) than the line display (mean=45.3 kts per trial). There was also a main effect of approach geometry [$F(2,1188)=3.703$, $p=.025$], which can best be explained in the context of the geometry by validity interaction [$F(2,1188)=9.183$, $p<.001$] which can be seen in figure 15, which depicts the RMS deviation per trial from a prescribed airspeed of 325 kts. The invalid predictor only had a cost to deviation on descending geometry trials where the mean deviation was 39.4 kts per trial compared to the valid/descending trials which had a mean deviation of only 31.2 kts per trial. The invalid predictor actually had a benefit on level approach geometries of 15 kts per trial.

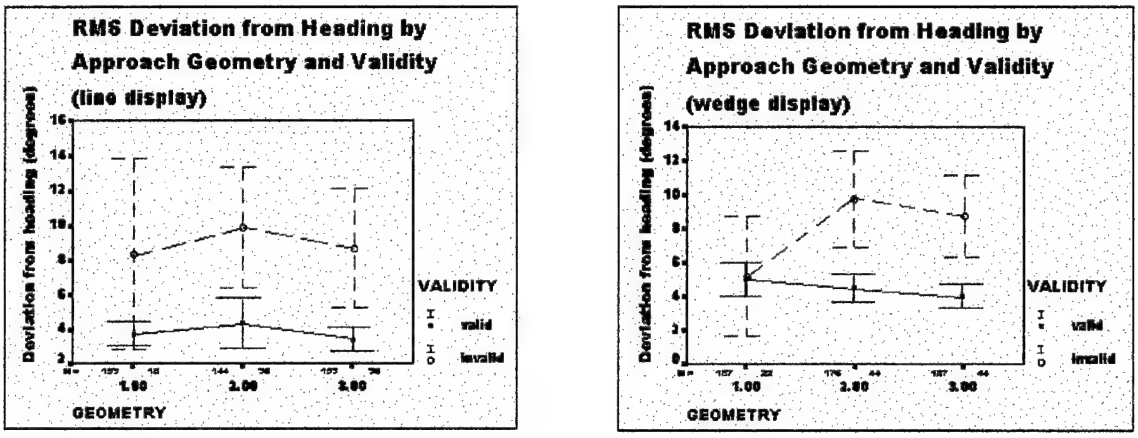


Figure 16: RMS deviation from heading measure (bar represents 95% confidence interval)

* * * A N A L Y S I S O F V A R I A N C E * * *

LOG(RMS deviation from Heading)

by VALIDITY DISPLAY GEOMETRY

UNIQUE sums of squares

All effects entered simultaneously

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig of F
Main Effects	16.323	4	4.081	10.185	.000
VALIDITY	9.920	1	9.920	24.759	.000
DISPLAY	1.788	1	1.788	4.463	.035
GEOMETRY	1.169	2	.585	1.459	.233
2-Way Interactions	4.160	5	.832	2.077	.066
VALIDITY DISPLAY	.290	1	.290	.724	.395
VALIDITY GEOMETRY	3.368	2	1.684	4.203	.015
DISPLAY GEOMETRY	.439	2	.220	.548	.578
3-Way Interactions	.456	2	.228	.569	.566
VALIDITY DISPLAY GEOMETRY	.456	2	.228	.569	.566
Explained	24.821	11	2.256	5.632	.000
Residual	475.984	1188	.401		
Total	500.806	1199	.418		

Table 5: ANOVA table of RMS error of heading

Figure 16 depicts the RMS heading deviation per trial from a prescribed heading of 360 degrees. Once again, this RMS deviation measure was analyzed using a logarithmic transformation to make the distribution appear more normal. Table 5, which is the ANOVA table of Log(RMS deviation from heading), shows that there were two main effects, that of predictor validity [$F(1,1188)=24.759$, $p<.001$], and of display type [$F(1,1188)=4.463$, $p=.035$]. The wedge display had a higher mean deviation (mean=5.22 degrees per trial) when compared to the line display (mean=4.73 degrees per trial). As expected, the invalid predictor lead to a higher deviation (mean=8.77 degrees per trial) compared to the valid predictor trials (mean=4.24 degrees per trial). The interaction between predictor validity and approach geometry [$F(2,1188)=4.203$, $p=.015$] can be seen in the right most points of figure 16, where the cost of the invalid predictor was exaggerated by the descending and level approach geometries. In the level geometry, this is a reversal of the effect seen in airspeed deviation.

3.1.3 Time to First Maneuver

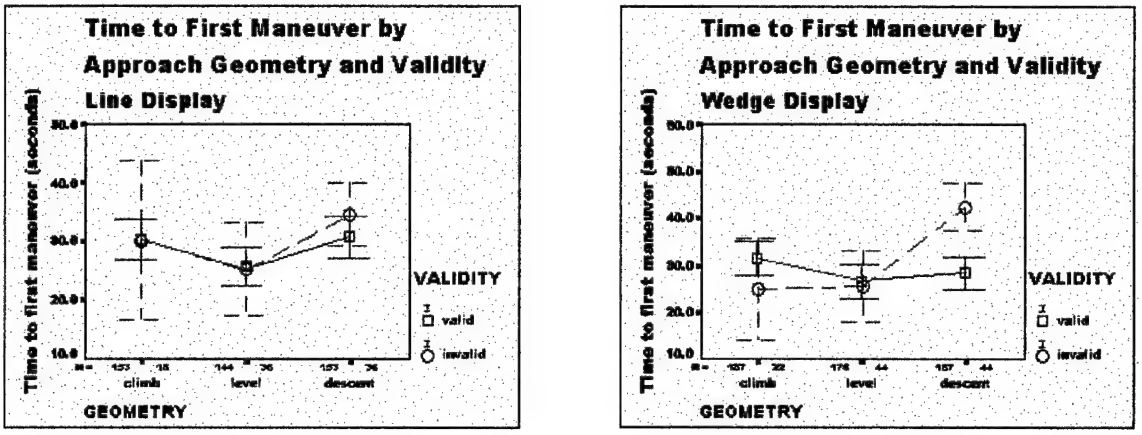


Figure 17: Time to first maneuver measure (bar represents 95% confidence interval)

Time to first maneuver

by VALIDITY DISPLAY GEOMETRY

UNIQUE sums of squares

All effects entered simultaneously

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig of F
Main Effects	93197758	4	23299439.602	4.492	.001
VALIDITY	4013014	1	4013014.381	.774	.379
DISPLAY	242397	1	242396.923	.047	.829
GEOMETRY	87667120	2	43833559.942	8.451	.000
2-Way Interactions	54818873	5	10963774.678	2.114	.061
VALIDITY DISPLAY	579088	1	579088.433	.112	.738
VALIDITY GEOMETRY	45340366	2	22670183.064	4.371	.013
DISPLAY GEOMETRY	5485558	2	2742778.760	.529	.589
3-Way Interactions	17839331	2	8919665.417	1.720	.180
VALIDITY DISPLAY GEOMETRY	17839331	2	8919665.417	1.720	.180
Explained	148800799	11	13527345.397	2.608	.003
Residual	6161901771	1188	5186786.002		
Total	6310702570	1199	5263304.896		

Table 6: ANOVA table of time to first maneuver

Figure 17 depicts the time between the start of a trial and the pilots' initial maneuver in seconds. The time to first maneuver measure had a main effect of approach geometry [$F(2,1188)=8.451$, $p=.001$] (see table 6), with the earliest maneuvers occurring with the level geometry (mean=26.1 seconds). There was also an interaction effect of validity and geometry [$F(2,1188)=4.371$, $p=.013$] that can be seen in figure 17. As with other dependent variables, the costs of an invalid predictor seem to be only manifest in the descending approach geometries. Descending/valid trials had a mean time to first maneuver of 29.5 seconds per trial, compared to descending/invalid trials, which had a mean time to first maneuver of 39 seconds per trial.

3.1.4 Type of First Maneuver

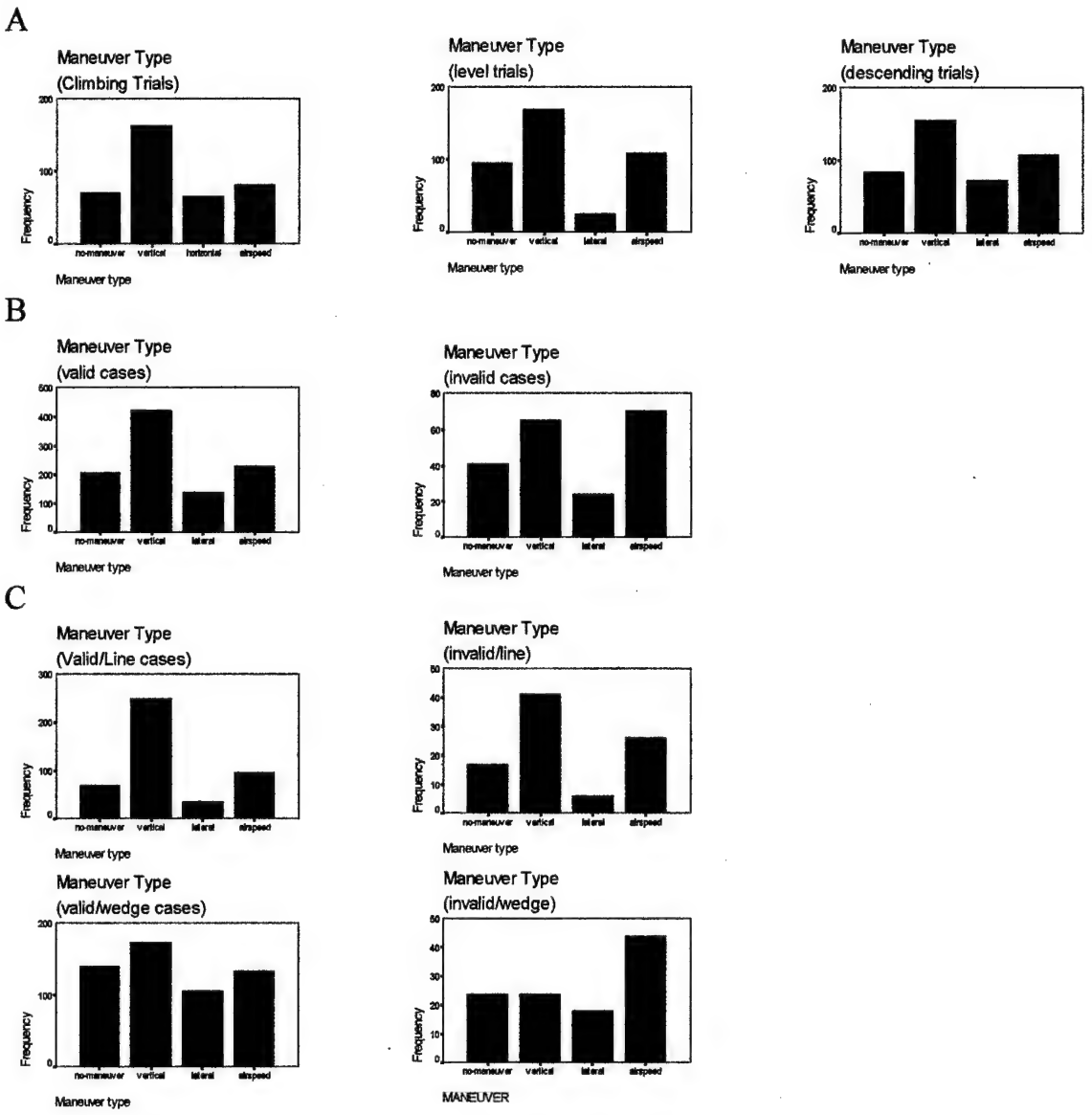


Figure 18: Type of first maneuver

Panel A: Averaged over validity and display for each approach geometry

Panel B: Averaged over approach geometry and display type for each level of validity

Panel C: Averaged over approach geometry for each level of validity and display

Figure 18 shows the type of first maneuver, broken down by whether the initial response was one of lateral, airspeed, or vertical control, and averaged across various dependent variables. Generally, vertical maneuvers were the most preferred, followed by airspeed and then heading changes. When the maneuver type was grouped by the three approach geometries (Row A), lateral maneuvers became more prevalent on climbing and descending trials than on level trials (center panel of row A). When grouped by the two levels of validity (Row B), we can see that on invalid trials (right panel), airspeed control is used with much greater relative frequency than on valid trials (left panel). Finally, when the maneuver type is shown broken down by the possible combinations of validity and display (C), we see that on invalid trials (right hand panels), the wedge display lead to a preference of airspeed maneuvers over vertical maneuvers, while the line display lead to the opposite preference.

3.2 Secondary Task Measures: FFOV Monitoring

3.2.1 Secondary Task Response Time

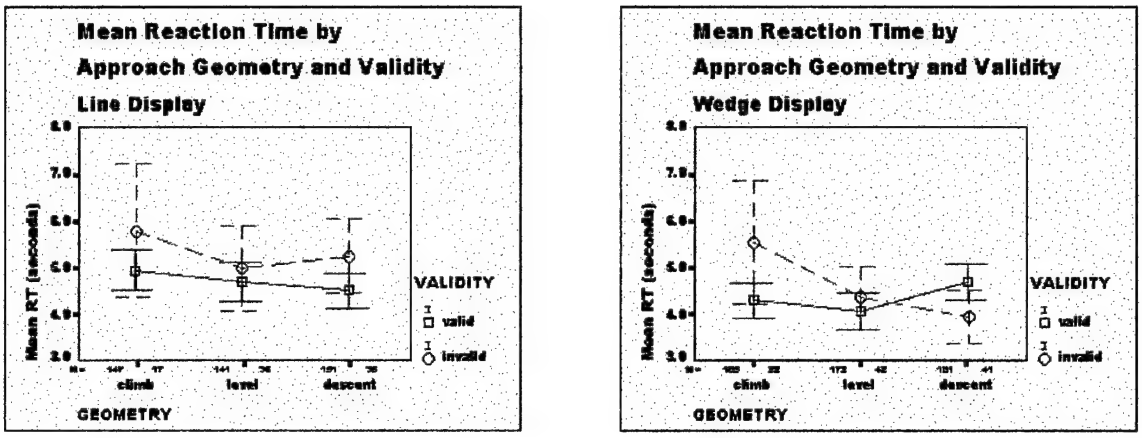


Figure 19: Mean reaction time to detection task measure (bar represents 95% confidence interval)

Log(Mean Response Time)
by VALIDITY DISPLAY GEOMETRY

UNIQUE sums of squares

All effects entered simultaneously

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig of F
Main Effects	.886	4	.221	3.232	.012
VALIDITY	.321	1	.321	4.683	.031
DISPLAY	.477	1	.477	6.964	.008
GEOMETRY	.274	2	.137	1.998	.136
2-Way Interactions	.196	5	.039	.574	.720
VALIDITY DISPLAY	.005	1	.005	.066	.797
VALIDITY GEOMETRY	.189	2	.094	1.378	.253
DISPLAY GEOMETRY	.003	2	.001	.019	.981
3-Way Interactions	.256	2	.128	1.869	.155
VALIDITY DISPLAY GEOMETRY	.256	2	.128	1.869	.155
Explained	1.828	11	.166	2.426	.006
Residual	79.344	1158	.069		
Total	81.172	1169	.069		

Table 7: ANOVA table of Log(response time to detection task)

Figure 19 depicts the response time to the FFOV indicators per trial. Response time had two main effects that can be seen in table 7, which shows the ANOVA table for the Log(RT) to the FFOV indicators. The first was of predictor validity [$F(1,1158)=4.683$, $p=.031$]. The invalid predictor lead to slower response times (mean=4.81 seconds) than the valid predictor (mean=4.52 seconds). The next effect was of display type [$F(1,1158)=6.964$, $p=.008$]. Here, the faster response time was with the wedge display (mean=4.37 seconds), and the slower response time with the line display (mean=4.80 seconds).

3.2.2 Secondary Task Accuracy

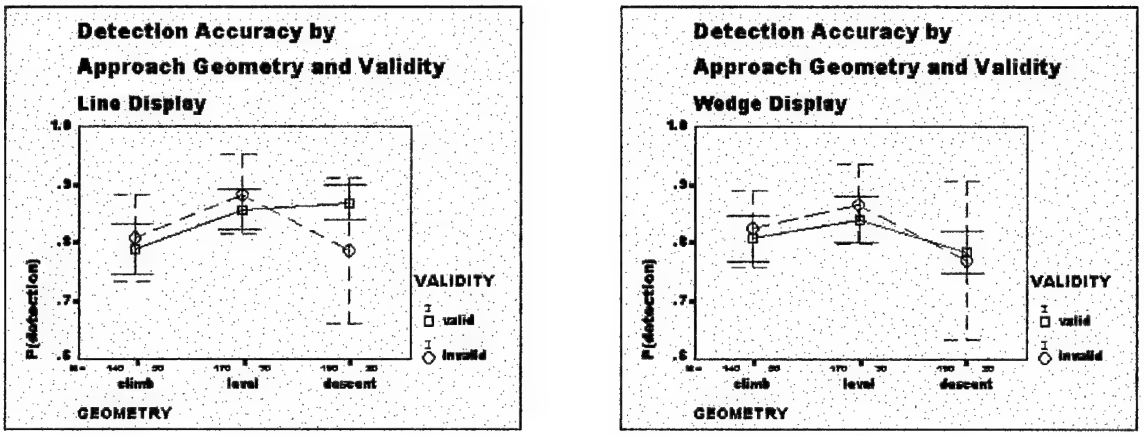


Figure 20: Accuracy of detection task measure (bar represents 95% confidence interval)

Detection Accuracy

by VALIDITY DISPLAY GEOMETRY

UNIQUE sums of squares

All effects entered simultaneously

Source of Variation	Sum of Squares	DF	Mean Square	F	Sig of F
Main Effects	.479	4	.120	2.023	.089
VALIDITY	.000	1	.000	.002	.960
DISPLAY	.041	1	.041	.695	.405
GEOMETRY	.437	2	.218	3.691	.025
2-Way Interactions	.320	5	.064	1.082	.369
VALIDITY DISPLAY	.018	1	.018	.309	.578
VALIDITY GEOMETRY	.135	2	.067	1.139	.320
DISPLAY GEOMETRY	.118	2	.059	.997	.369
3-Way Interactions	.033	2	.017	.281	.755
VALIDITY DISPLAY GEOMETRY	.033	2	.017	.281	.755
Explained	1.372	11	.125	2.107	.017
Residual	70.312	1188	.059		
Total	71.684	1199	.060		

Table 8: ANOVA table of accuracy of detection task

For response accuracy, P(detection), there was a main effect of approach geometry [$F(2,1188)=3.691$, $p=.025$] that can be seen in table 8. Figure 20 depicts the percent accuracy, indicating the greatest accuracy for level trials. The false alarm rate was very low in this measure (fewer than 5 false alarms throughout the 1200 non-practice trials), and was not a significant factor in the accuracy analysis. The accuracy between the three geometries was 80% for climbing, 82% for descending, and 85% for level.

3.3 Qualitative Analysis of Trust

The analysis of trust was done in a qualitative method by examining increases and decreases of safety measures, response times to the secondary task, and time to first maneuver between the trial immediately prior to an invalid trial and immediately following an invalid trial. In this case, we only employed the invalid trials that made the maneuvering more difficult by decreasing time to loss of separation (see figure 8 in section 2.3).

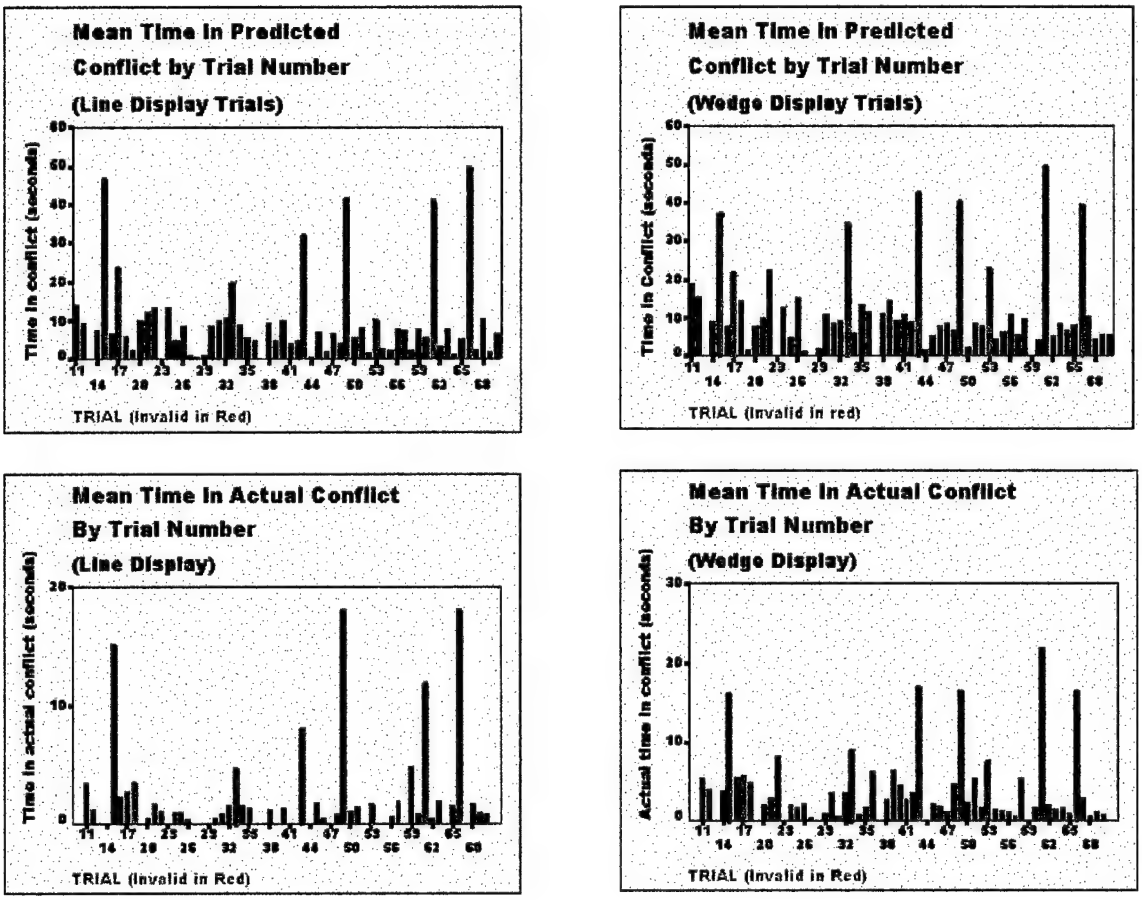


Figure 21: Mean time in predicted conflict and actual conflict for each non-practice trial, where invalid predictors that lead to less time to loss of separation trials are highlighted in red

As shown in Figure 21, the times in conflict are inflated during invalid trials (red bars), which was analyzed previously in the context of figures 12 and 13. The time pilots spent in predicted conflicts (top two panels) tended to decrease from the trial immediately prior to an invalid trial to the trial immediately following the invalid trial (11/16 occurrences), and the trend was consistent across both line and wedge displays (6/8, and 5/8 occurrences respectively). For time in actual conflicts (bottom two panels), there was no

real trend to increase or decrease time in conflict following an invalid trial. None of the trends were very strong, and were not statistically significant.

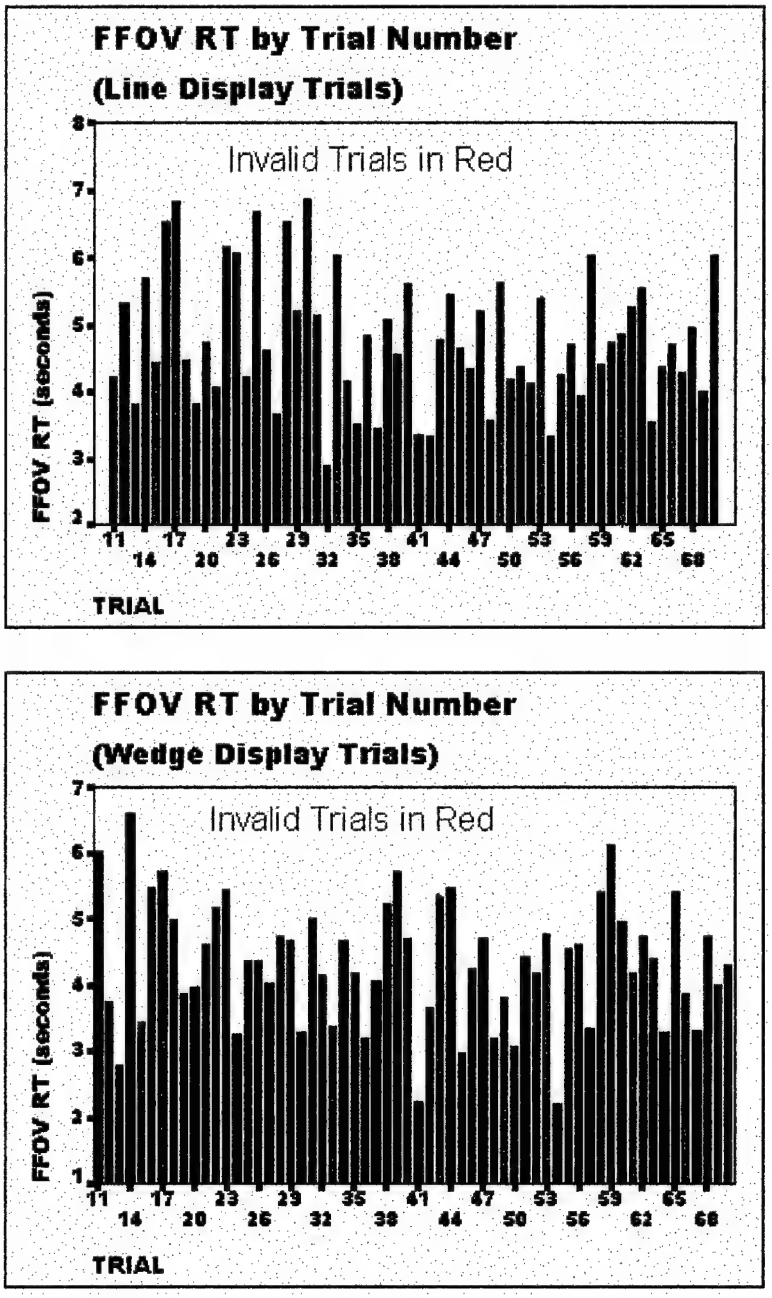
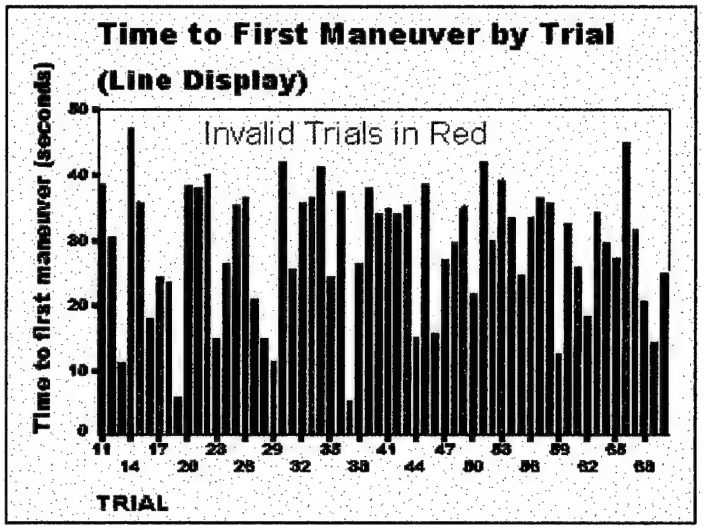


Figure 22: Mean response time to FFOV indicators for each non-practice trial, where invalid predictors that lead to less time to loss of separation trials are highlighted in red

As seen in figure 22, in response time to the FFOV indicators, there was a trend to increase RT from before the invalid trial to after the invalid trial when using the line display (5/8 occurrences). This trend is reversed in the case of the wedge display (6/8 occurrences). However, both trends are extremely weak, and probably do not suggest any substantial change in trust (as manifest by the secondary task) following an invalid trial.



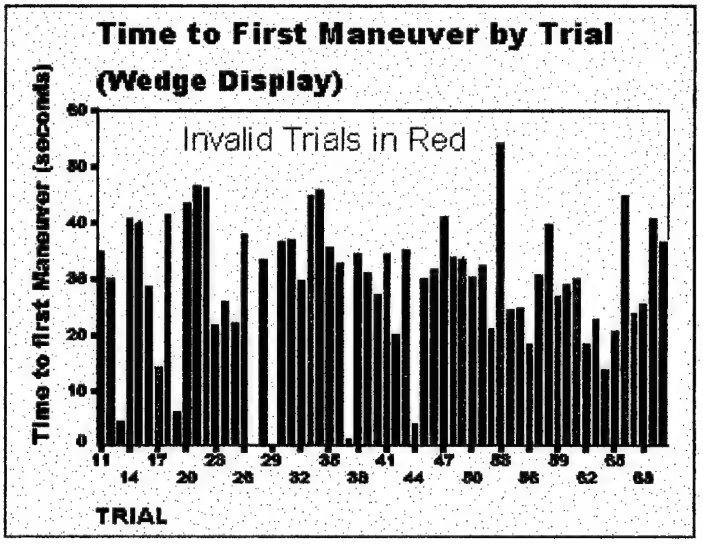


Figure 23: Time to first maneuver for each non-practice trial, where invalid predictors that lead to less time to loss of separation trials are highlighted in red

Figure 23 represents the mean time to first maneuver for each trial. There was no trend of either increasing or decreasing time to first maneuver between the trial prior to an invalid trial and the trial following an invalid trial for the line or wedge display. Exactly ½ of the occurrences increased and ½ decreased time to first maneuver after an invalid trial occurred. Thus, in general there does not appear to be any behavioral manifestation of trust shifts brought on by an invalid trial in these particular dependent measures.

3.4 Subjective Data

Subjective data were collected by a post-experiment questionnaire where subjects were asked to state their preferred type of maneuver to avoid conflicts if they had one. They were also asked to estimate what number of trials of the 60 non-practice trials had an invalid predictor. Since the display type was a between subjects variable, we could not gather information about pilots' preference of one display over another.

A B C

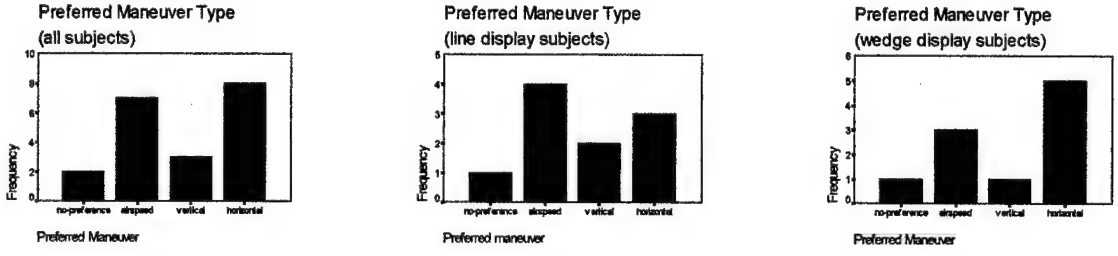


Figure 24: Pilot reported preferred type of maneuver Panel A: For all subjects Panel B: For subjects using the line display Panel C: For subjects using the wedge display

Figure 24 shows the pilots' preferred type of maneuver to avoid conflicts broken down by display type, which shows that there was the same trend for preferred maneuver type for subjects who used the line display as for subjects who used the wedge display, although the wedge display pilots tended to prefer horizontal maneuvers at a higher relative frequency than did the line display users. These data do not seem to correspond with what maneuvers were actually taken. Pilots reported that they preferred airspeed and horizontal maneuvering more than vertical maneuvering. However, the collected data showed that they were more likely to make vertical and airspeed maneuvers than horizontal maneuvers as seen in figure 18. However, since the data collected for figure 18 reflect only the initial maneuver, the data may not perfectly correspond to that collected for preferred maneuver type.

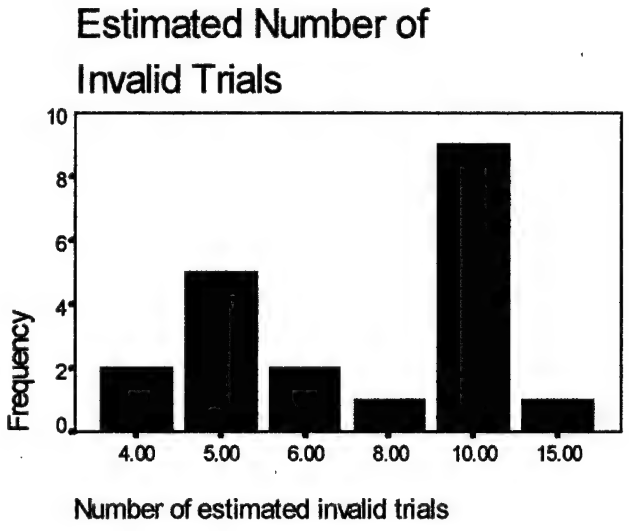


Figure 25: Pilot's estimate of the number of invalid trials over all 60 trials

Figure 25 shows a histogram of the frequency of pilots' different estimates of the number of invalid trials in the experiment. From Figure 25, we can tell that pilots were reasonably accurate in their estimate of how many of the trials contained an invalid predictor. In actuality, 10 of the trials had an invalid predictor, of which 8 would require more maneuvering. The mean response was 7.9 invalid trials.

4.0 Discussion

Examining the results as a whole, the most prevalent findings across several dependent variables were those of approach geometry, predictor validity, and their interactions. In general, several dependent variables revealed that intruder descending trials were the most difficult (see figures 12, 13, 14, and 17). This finding is contrary to an effect found by Wickens and Morphew (1997), who found that descending traffic produced fewer predicted conflicts than level or climbing traffic. However, it is consistent with the findings of Merwin and Wickens (1996), who found that climbing and descending geometries were the most difficult, though the coplanar display format minimized this geometry cost, and consistent with the results of the current experiment, Merwin and Wickens also found that the level trials were the easiest.

Since our subjects tended to maneuver vertically most often (see figure 18), this cost of vertical approach geometries can be explained by the fact that vertical approaches required more cognitive resources to avoid, in that they involved two axes of approach. The intruder would be approaching both laterally and vertically, while on level approaches, there was only the lateral axis of approach. Therefore, vertical approaches required the pilot to make two decisions: 1) whether or not the intruder would close to within three miles horizontally and 2) whether or not the intruder would come within 1500 ft once within that three mile ring. However, during level approaches, the pilot could immediately ascertain if the intruder was within 1500 ft and only had to make

one decision: whether or not the intruder would come within three miles. Since pilots tended to maneuver more vertically (see figure 18), they also had to contend with two axes of deviation from their prescribed trajectory, both altitude and airspeed, since airspeed changed rapidly with climbs and descents without constant correction (with the joystick buttons) of airspeed to 325 kts. Each of these factors (more decisions and more control responses required), corresponding to vertical approach geometries, would make them more difficult. The fact that intruder descending trials were more difficult than climbing trials in the current experiment is as yet unclear.

As expected, several dependent variables showed that performance suffered on trials with an invalid predictor, either across the board and/or in terms of the interaction with geometry (see figures 12, 13, 14, 15, 16, and 17). This form of interaction was almost always one where invalidity cost was amplified (or shown exclusively) on the more difficult descending trials (see figures 12, 13, 14, 15, and 17), the only exception being that of RMS deviation from heading, which showed invalidity cost to be highest on level geometries.

Since the intruder descending trials seemed to be the most difficult from the previous discussion, such trials presumably required the most cognitive resources in terms of visualization of the conflict, therefore pilots needed to rely to a greater extent on the predictive display. This over-reliance on the invalid predictor is likely the same "garden path" effect found by Conejo and Wickens (1997) and Yeh, Wickens, and Seagull (1998), who found that subjects tended to rely on automated cueing regardless of its reliability. Also,

Kirschenbaum and Arruda (1994), who used a graphical ellipse display to show the confidence in a naval target's position, found benefits for their reliability display only on more difficult trials. This suggests that the uncertainty cost mediated by the display was only present on difficult trials, which leads to a discussion of the unreliability (i.e. wedge) display itself.

The primary thrust of this experiment was to ascertain if any invalidity cost could be minimized through an explicit display of predictor reliability without increasing head-down time, or hurting safety measures. If indeed the wedge display did help attenuate the costs of unreliability, as was predicted, we would expect to have had some significant 3-way interactions. Specifically, the descending geometry unreliability cost (the two-way interaction seen in so many of the dependent variables) would have been attenuated by the wedge display relative to the line display (a 3-way interaction). In fact, no such 3-way interactions were found, and in no dependent measure were any of the 3-way interactions at all close to significance in a direction that might indicate such attenuation. Indeed the costs of following the invalid predictor corresponded completely with the findings of Conejo and Wickens (1997), and Yeh, Wickens, and Seagull (1998), who found "garden path" effects with displays that did not show their reliability. Therefore, it appears that the wedge display did not contribute to the pilots' trust calibration and perceived reliability as depicted in figure 4 of the introduction, but instead the wedge display had some more enduring effects across all trials regardless of validity and geometry (see tables 2, 4, 5, and 8). Some of these effects were counterproductive (see figure 16).

The wedge had a cost to the safety measures of time spent in actual conflicts, and efficiency in deviation from heading. The magnitude of the cost to time in actual conflict was 2 seconds per trial (48%), and of heading deviation .49 degrees per trial (9.4%).

The wedge also had a benefit to secondary task performance (.43 second shorter response time), an effect of 9%. Incidentally, the mean RT in this experiment corresponded with that found by Morphew and Wickens (1996) (mean RT=4.6 seconds and 4.1 seconds respectively). The wedge also had a significant benefit to RMS deviation from airspeed of 16 kts per trial (and effect of 35%). This benefit to airspeed deviation is much greater than its cost to heading deviation (35% benefit, 9.4% cost). The wedge display response time benefit tends to suggest the presence of the wedge led to a shift in attention allocation strategy away from the primary traffic display, towards the FFOV, thus shortening response time to events in the FFOV but being slightly less vigilant on monitoring the traffic display, leading to the 1.8 second per trial increase in time in conflict.

The previously mentioned costs of invalid predictors suggests trust and complacency problems like the "garden path" effect found by Conejo and Wickens (1997). However, no real effects were found in comparing trials immediately before to trials immediately after a failure in terms of their time in conflict or response times to the secondary task. This suggests that no real trust changes occurred due to the invalid trials, but that the invalid trials themselves reduced performance dramatically, as seen in figures 21 and 22.

This is consistent with the findings of Lee and Moray (1992), who found that performance recovered quite quickly after an automation failure, though subjective ratings of trust took several trials to approach previous levels. In fact, this may be a positive finding for the CDTI predictor itself, because the automation failure did not have long lasting safety effects, but only lead to performance costs on the trials where the predictor fails.

The subjective data show that the maneuvers pilots preferred, were not necessarily the maneuvers that they made to avoid traffic conflicts. It may be that while pilots may prefer one type of a maneuver (horizontal changes) over another (vertical changes), they may be forced by the scenario to perform the less desirable maneuver. An alternative explanation is also that these two analysis actually measure different things. Pilots may not perform their preferred maneuver initially on each trial, and the objective data of maneuver choice used in this experiment were only those of the maneuver initiated first.

In conclusion, our results showed that pilots can still become over-reliant on useful automation tools in general, regardless of the tool's reliability. Specifically, pilots are more reliant on these tools as the cognitive cost of not using the tool increases, as in the case of the descending approach geometry condition in the current experiment. This particular attempt to better calibrate trust in the automation tools was unsuccessful, though other methods of displaying reliability may be more effective in the CDTI, such as the luminance reliability display used by Montgomery and Sorkin (1996), or a specific limited number of levels of reliability being displayed, like that used in the likelihood

alarms of Sorkin, Kantowitz, and Kantowitz (1988). Ideally, a more appropriate display of reliability that will facilitate pilots' trust calibration still needs to be found to make the CDTI predictor more useful in real world application like free-flight, knowing that the predictor *can not* be perfectly reliable. In the case of predictor reliability, methods of ascertaining subjective values of trust should be employed, like those used by Lee and Moray (1992), and Muir and Moray (1996). Future research in this area should focus on that search for promoting proper trust calibration in automation tools in general.

References

Barhydt, R., and Hansman, R.J. (1997). Experimental studies of intent information on cockpit traffic displays. 9th International Symposium on Aviation Psychology.

Conejo, R., and Wickens, C.D. (1997). The effects of highlighting validity and feature type on air-to-ground target acquisition performance. . University of Illinois Institute of Aviation Technical Report (ARL-97-11/NAWC-ONR-97-1). Savoy, IL: Aviation Res. Lab.

Ellis, S.R., McGreevy, M.W., and Hitchcock, R.J. (1987). Perspective traffic display format and airline pilot traffic avoidance. Human Factors 29(4). 371-382.

Gallaher, P.D., Hunt, R.A., and Willeges, R.C. (1977). A regression approach to generate aircraft predictor information. Human Factors, 19(6), 549-555.

Hart, S.G., and Wempe, T.E. (1979). Cockpit display of traffic information: Airline pilot's opinions about content, symbology, and format. NASA Technical Memorandum 78601. Moffet Field, CA: Ames Res. Center.

Jago, S., and Palmer, E. (1982). Separation monitoring with four types of predictors on a cockpit display of traffic information. 16th Ann. Conf. On Manual Control. 439-446.

Jensen, R.S. (1981). Prediction and quickening in perspective flight displays for curved landing approaches. Human Factors. 23(3), 355-363.

Johnson, W.W., and Battiste, V. (1997). Development and demonstration of a

prototype free flight cockpit display of traffic information. Proceedings of the 1997 SAE/AIAA World Aviation Congress.

Kantowitz, B.H., Hanowski, R.J., & Kantowitz, S.C. (1997). Driver acceptance of unreliable traffic information in familiar and unfamiliar settings. Human Factors. 39(2), 164-176.

Kirschenbaum, S.S., & Arruda, J.E. (1994). Effects of graphic and verbal probability information on command decision making. Human Factors, 36(3), 406-418.

Kreifeldt, J.G. (1980). Cockpit displayed traffic information and distributed management in air traffic control. Human Factors, 22(6), 671-691.

Lee, J. & Moray, N. (1992). Trust, control strategies and allocation of function in human machine systems. Ergonomics. 35(10), 1243-1270.

Lee J. & Moray, N. (1994). Trust, self confidence, and operators' adaptation to automation. Int. J. Human-Computer Studies. 40, 153-184.

Levison, W.H., (1982). The optimal control model for the human operator: Theory, validation, and application. Proceedings of the workshop on flight testing to identify pilot workload and pilot dynamics. Edwards AFB, CA.

Montgomery, D.A., and Sorkin, R.D. (1996). Observer sensitivity to element reliability in a multi-element visual display. Human Factors, 38(3), 484-494.

Muir, Bonnie M, (1987). Trust between humans and machines, and the

design of decision aids. Special Issue: Cognitive engineering in dynamic worlds. International Journal of Man-Machine Studies. 27(5-6) 527-539.

Muir, B.M., and Moray, N. (1996). Trust in Automation. Part II. Experimental studies of trust and human intervention in a process control simulation. Ergonomics. 39(3), 429-460.

National Transportation Safety Board. (1973). Eastern Airlines L-1011, Miami, Florida, 29 Dec. 1972 (Report no. NTSB-AAR-73-14). Washington, D.C.

O'Brien, J.V. and Wickens, C.W. (1997). Free flight cockpit displays of traffic and weather: Effects of dimensionality and data base integration. Proceedings of the 41st Annual Meeting of the Human Factors and Ergonomics Society. Santa Monica, CA: Human Factors and Ergonomics.

Palmer, E. (1983). Perception of horizontal aircraft separation on a cockpit display of traffic information. Human Factors Journal, 22(5), 605-620.

Palmer, E., Jago, S., & Dubord, M. (1981). Horizontal conflict resolution maneuvers with a cockpit display of traffic information.

Parasuraman, R., and Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors 39(2) 230-253.

Poulton, E.C. (1957). On prediction in skilled movements. Psychological Bulletin. 54(6) 467-478.

RTCA, (1995). Free flight implementation. RTCA Task Force 3 Report. Washington D.C.: RTCA Inc.

Singh, I.L., Malloy, R. and Parasuraman, R. (1993). Individual

differences in monitoring failures of automation. The Journal of General Psychology 120(3), 357-373.

Sorkin, R.D., Kantowitz, B.H., and Kantowitz, S.C. (1988). Likelihood alarm displays. Human Factors, 30(4), 445-459.

Sparaco, P. (1994). A330 crash to spur changes at Airbus. Aviation Week and Space Technology, 141(6), 20-22.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124-1131.

Wickens, C.W. (1986). The Effects of control dynamics on performance. In K.R. Boff, L. Kaufman, J.P. Thomas (Eds.), Handbook of Perception and Human performance (Volume II) (pp. 39.1-39.60). New York: John Wiley & Sons. 1986.

Wickens, C.D., and Carswell, C.M. (1997). Information processing. In G. Salvendy (ed.), Handbook of Human Factors & Ergonomics. (2nd ed), New York: John Wiley. 1997.

Wickens, C.D., Gordon, S.E., & Liu, Y. (1998). An Introduction to Human Factors Engineering. New York, NY: Addison Wesley Longman, Inc.

Wickens, C.D., Mavor, A.S., Parasuraman, R., & McGee, P. (1998) Airspace system integration: the concept of Free Flight. The Future of Air Traffic Control. Washington D.C.: National Academy Press. 1998.

Wickens, C.D., & Morphew, E. (1997). Predictive features of a cockpit traffic

display: A workload assessment. University of Illinois Institute of Aviation Technical Report (ARL-97-6/NASA-97-3). Savoy, IL: Aviation Res. Lab.

Wickens, Pringle, and Merlo (in preparation). Attention, reliability and trust in information integration: implications for display design. Aviation research Lab/Army Research Lab technical report: Oct, 1998.

Yeh M., Wickens, C.D., & Seagull, F.J. (1998). Effects of frame of reference and viewing condition on attentional issues with helmet mounted displays. University of Illinois Institute of Aviation Technical Report (ARL-98-1/ARMY-FED-LAB-98-1). Savoy, IL: Aviation Res. Lab.